

•

**TESTS AND MEASUREMENTS IN
ELEMENTARY EDUCATION**

9

Tests and
Measurements
in
Elementary Education

M. J. Nelson, *Dean of the Faculty*
Iowa State Teachers College

THE CORDON COMPANY
NEW YORK

Copyright, 1939, by
The Corson Company, Inc.
225 Lafayette Street
New York, N. Y.

ALL RIGHTS RESERVED. NO PART OF THIS BOOK MAY
BE REPRODUCED IN ANY FORM, BY MIMEOGRAPH OR
ANY OTHER MEANS, WITHOUT PERMISSION IN WRIT-
ING FROM THE PUBLISHER, EXCEPT BY A REVIEWER
WHO WISHES TO QUOTE BRIEF PASSAGES AS PART OF
A REVIEW IN A MAGAZINE OR NEWSPAPER

MANUFACTURED IN THE UNITED STATES OF AMERICA

TO
THE MEMORY
OF
FRANK L. CLAPP

PREFACE

THE WRITER of a present-day textbook in the field of testing must take cognizance of the fact that developments in the field are taking place in two major directions. In the first place, there is much consideration being given to the improvement of standardized tests by both test makers and publishers, and a very distinct effort to make such tests yield more usable information concerning the outcomes of instruction. In the second place, there is greatly increased attention to the improvement of tests made by the classroom teachers and to the use of such tests for instructional purposes.

In dealing with standardized tests it was found wholly impossible, even if it had been thought advisable to do so, to attempt to mention all of the standardized tests now being published. It has seemed best, therefore, to discuss only a few tests representative of those available. In choosing particular tests for mention, preference has been given to those that point out certain desirable aspects of instruction or that serve to illustrate procedures and problems which the teacher should have in mind. It is hoped, therefore, that this volume will be of some value to the classroom teacher as well as to the inexperienced teacher who is beginning a serious study of the problems of education.

Instructors who use this volume as a textbook may wish, in taking up the various topics, to depart from the order in which they are here presented. For example, some may wish to present the criteria for the selection of tests, before the discussion of standardized tests themselves. This may be accomplished by taking up Chapter XII immedi-

ately following Chapter IV. Others may wish to take up the uses of tests as discussed in Chapter XIII before a consideration of tests in particular fields, for example, following immediately after the discussion of test construction in Chapter III. Still others may prefer to have the discussion of intelligence testing (Chapter XI) precede that of achievement testing (Chapters V-X). These changes may be made without disturbing the continuity of treatment.

This book was originally projected as a co-operative enterprise of the late Professor Frank L. Clapp of the University of Wisconsin and the present writer. Since very little of the actual writing had been done prior to Professor Clapp's untimely death, and since the organization of the material has been almost completely altered from the original plan, the writer must accept the sole responsibility. He does, however, wish to acknowledge his indebtedness to his late co-worker for valuable suggestions and inspiration.

To many others the writer is deeply grateful for valuable assistance and advice. He desires especially to mention Miss Gladycé Gooder, who so carefully prepared the manuscript for the printer. Among his colleagues at Iowa State Teachers College, the most helpful have been Dr. J. B. Paul, Director of Research, who read the manuscript and offered helpful suggestions, and Dr. A. E. Brown, Professor of Education, and Dr. E. C. Denny, Head of the Department of Education, who used the text in their classes in mimeographed form and offered practical criticism. Some of the writer's own students also contributed useful suggestions. Finally, the writer expresses appreciation to the many authors and publishers who have granted permission to quote from their copyrighted works.

April, 1939

M. J. NELSON

TABLE OF CONTENTS

<i>CHAPTER I</i>	<i>PAGE</i>
The General Character of Measurement in Education	15
 <i>CHAPTER II</i>	
Classification of Tests	37
Classification as to Function	38
Classification as to Form	45
 <i>CHAPTER III</i>	
Constructing and Scoring Tests	59
 <i>CHAPTER IV</i>	
Common Statistical Terms and Procedures	83
 <i>CHAPTER V</i>	
Tests in Reading and Arithmetic	109
Tests in Reading	109
Tests in Arithmetic	120
 <i>CHAPTER VI</i>	
Tests in Language, Handwriting, and Spelling	139
Tests in Language	139
Tests in Handwriting	150
Tests in Spelling	155
 <i>CHAPTER VII</i>	
Tests in the Social Studies	167
History Tests	173
Civics Tests	174
Geography Tests	174

<i>CHAPTER VIII</i>	<i>PAGE</i>
Tests in Music and Art	185
Tests in Music	185
Tests in Art	195
 <i>CHAPTER IX</i>	
Miscellaneous Tests for Elementary Schools . . .	205
Tests in Health Education	205
Tests in Elementary Science	211
Personality and Character Tests	214
 <i>CHAPTER X</i>	
General Tests of Achievement.	231
 <i>CHAPTER XI</i>	
The Measurement of Intelligence	251
 <i>CHAPTER XII</i>	
Criteria for the Selection of Tests	275
 <i>CHAPTER XIII</i>	
Using Educational Tests	301
 <i>CHAPTER XIV</i>	
Recording and Reporting Progress	323
 Appendix	341
 Index	345

•

**TESTS AND MEASUREMENTS IN
ELEMENTARY EDUCATION**

•

•

**THE GENERAL CHARACTER OF
MEASUREMENT IN EDUCATION**

•

•

Chapter I

THE GENERAL CHARACTER OF MEASUREMENT IN EDUCATION

Main *The Importance of Measurement in General*

MEASUREMENT IS SO COMMON in the experiences of those of us who are living now that it is difficult for us to think of a time when there were no exact procedures for measuring distance, weight, time, area, volume or, in fact, any of the many values that we now measure with great accuracy. Also, it is difficult for us to appreciate the value that our present systems of measurement have, and to what extent many of the comforts and conveniences in our lives depend upon accurate measurement.

To illustrate the extent to which we depend on measurement now we may think of Mr. A. He lives five miles from the railroad station. He knows the distance because it has been *measured*. He knows when to leave his home in order to reach the station in time to take his train because his watch *measures* time. The speedometer on his car *measures* the speed at which he drives and also how much of the distance he has covered at any particular time. When he buys his ticket he pays for it with money, the value of which is *measured* in exact units. In his office he finds the supply of heat controlled by an instrument which *measures* the temperature. At his club he finds the

lunch delightful because the chef has carefully *measured* the various ingredients of the dishes he has prepared. The newspaper that he reads after his lunch has lines of uniform length, well-balanced headings, and pleasingly arranged advertisements all made possible by *measurement*.

We can see the value of measurement in such connections as have just been cited, but it is not so easy to realize that the advance of civilization has depended largely on the development of accurate systems of measurement. The measurement of time, in terms of years, months, and days, has made possible accurate records of nations, definite treaties, movements of armies, orderly government; the measurement of distance, weight, and volume has made possible the construction of roads, railways, canals, ships; measurement has played a great part in the development of architecture, chemistry, physics, astronomy, medicine, commerce, and agriculture. In fact, history makes it clear that no group of people has advanced very far in civilization until it has developed systems of measurement.

Any one of the special fields named in the preceding paragraph was aided greatly in its own development and, consequently, in its contribution to the advance of society, by the development within the field of some method of measuring conditions peculiar to that field. For example, physics was aided by the development of methods of measuring the intensity of light, the amount of force or power, the amount of electricity; medicine profited greatly by the invention of methods of measuring the temperature of the human body and blood pressure; agriculture benefited when methods were devised for measuring the quality of soil, the quality of seed, the various contents of milk.

It seems reasonable to suppose that education is not

entirely unlike other fields, but that the perfecting of a system which would make it possible to measure accurately the results of instruction would be of great assistance in the development of education, and in the contribution that education may make to the progress of society.

The system of measurement in education as developed up to the present time is far from perfect and there is still a great amount of research work necessary before the measuring instruments in education can approach the accuracy of measurement in other fields. The student of educational measurement must realize that he is studying a movement that is new, not only to him, but to all educators,—a movement that, we might say, is in its infancy. Since this is true it would seem proper at this point to give some attention to the way in which systems of measurement in other fields were developed, and later to compare the procedure used in their development with that now being used in developing systems of measurement in education.

The Development of Systems of Measurement in General

A brief description of how one system of measurement began and developed will perhaps help to make clear how systems of measurement in general have been developed. For illustration we will take our system of linear measurement—the measurement of distance.

The reader must go back in his thinking to a time when there were no such conceptions in the minds of men as those represented by the terms “mile,” “rod,” “yard,” “foot,” “inch,”—no conception of any unit of length with a fixed value.

As it became advantageous to express some notion of length or of distance, men began to make comparisons between the length to be described and something with which everybody was familiar. Various parts of the human body were chosen, the forearm among the first. In practice this "unit" of length was defined as the bent forearm from the elbow point to the finger tip. A certain length was spoken of as so many "forearms." In time this "unit" was spoken of as a "cubit" and things were described as so many "cubits" long, or high. Other "units" employed were the "span" (the distance that could be reached by spreading the thumb and the little finger as far apart as possible), the "palm" (the breadth of the four fingers), and the "digit" (the breadth of the middle finger). Obviously not all forearms were of the same length so that a "cubit" was not a unit of measure with a fixed or definite value, and no more were the other "units."

The next step was to fix a definite value for the units employed. This was done in the case of the cubit by taking the typical, normal, or average forearm. Thus the cubit came to have a value now expressed as 18.24 inches. In most nations this was the value of the cubit as early as 4000 B.C. From this time on the development of linear measurement was simply a matter of agreeing upon the size of units, large and small.

Various nations now employ different units of measurement. In 1875, nineteen countries, including the United States, signed an agreement to establish an International Bureau of Weights and Measures to be located near Paris. This bureau was established and to it from time to time are referred the standards of measurement used in the various nations. The Metric System has been officially adopted in many countries, but its use has not become

common in all countries that have adopted it. It is, however, used extensively in international affairs.

Students of the history of measurement say that no one of the common units of measurement is entirely arbitrary but that each has some reasonable basis in human experience. The preceding paragraph makes it clear that this was the case with the earliest units of length, the cubit, the span, and others having their origin in the human body. The cubit of ancient Egypt and the various common units of linear measurement of today, such as the foot, yard, meter, mile, probably had their present values determined indirectly by reference to the distance on the earth's surface between the pole and the equator.

Measurement of Quantity

Nearly all systems of measurement are means of measuring quantity. We have a system for measuring distance or length, a system for measuring area, a system for measuring volume, a system for measuring weight, and numerous other systems for measuring values that are quantitative. By "quantitative" is meant something that exists in the physical world, having some such qualities as size, dimension, and content. It can usually be seen, and its presence or absence is easily noted.

Present-day systems for measuring quantity are made up of units each having a fixed value and with a constant ratio between one and another. For example, in our system of measuring length we have a table which reads as follows:

12 in.	= 1 ft.
3 ft.	= 1 yd.
5½ yds.	= 1 rd.
320 rds.	= 1 mi.

Each of the units in this table are referable directly to zero, that is, to a total absence of length. This means, for example, that when one is told that a stick is 3 feet long, he has a definite idea of the length of the stick—he does not need to refer “3 feet” to anything other than zero.

Thus we see that one essential in a system of measurement which is satisfactory, is that each “unit” connotes a certain definite value above “o.” Another essential is that on instruments which represent the system of measurement, all units bearing the same name must have the same value. For example, all the inches on a foot rule must be of the same length; all the feet on a yardstick must be of the same length; all pint containers must have the same volume; all pound weights must weigh the same.

In measuring quantity, the measurement is complete. For example, we weigh *all* of the sugar that we buy; we measure the length of the entire piece of board that we want to saw off; we measure the entire capacity of the grain bin. In actually measuring the length of a board we lay our measuring instrument (a yardstick, perhaps) on the board—we compare directly the length of the board with the standard units (yard, foot, inch) of length on the yardstick. Such measurement is not a matter of opinion or of subjective estimate; it is the result of applying to a quantitative situation a fixed system of measurement which is characterized by a point “o,” and by definite units on a scale. Such measurement, we say, is objective. It is quite different from the measurement of quality as described below.

Measurement of Quality

In connection with the measurement of quantity we described something so common in our lives as to make a

description seem quite unnecessary and almost ridiculous. The purpose of that description was to contrast with the very definite measurement of quantity the very indefinite measurement of quality. As a rule, measurement of quality, although quite common in our experience, is by no means as satisfactory as is our measurement of quantity. This will be obvious if we think of a few situations in connection with which we use qualitative measurement. We measure (judge) the quality of clothing; we measure (judge) the quality of a car; we measure (judge) the quality of a horse; we measure (judge) the quality of the work which teachers do.

When one measures the quality of the cloth in a suit of clothes, he has in mind some "standard" to which, for example, cloth that is described as "wool" should conform. This "standard" exists only in the minds of individuals—it is expressed in nothing comparable to a yardstick. The standard with which a person compares the cloth in the suit may be a very good standard or a very poor one; that is, his ideas of woollen goods may or may not be correct. The impression that he gets of the material in the suit may be dependable or it may be quite undependable. In any case, the measurement is largely a subjective procedure. Whatever test a person may apply to the material in clothing he does not attempt to apply his measurement to the entire article but he is satisfied to examine only parts of it.

In measuring the quality of work done by a teacher, a principal or superintendent does not see all of the work of the teacher; he simply visits an occasional class or, perhaps, from time to time, he spends as much as a half-day observing the teacher. In this way he takes what we may call "samples" of the work of the teacher. Judging from

these samples he rates the teacher as "superior," "good," "fair," or "poor." When he rates the teacher as "good" he does not say how far she is from having no value as a teacher, as is the case when we say a quantity of apples weighs twenty-five pounds. In the latter case, the quantity of apples is described as being made up of twenty-five units, each having the same measureable value, namely, one pound. In the case of the teacher who is rated "good," the term "good" (which corresponds to the quantitative measurement of twenty-five pounds in the case of the apples) has no definite value—there are no "units" of a definite value which make up a "good" teacher. A teacher is described as "good" when she is compared in a general way with other teachers whose work the superintendent has known. One person may rate a teacher as "good" while another person may rate the same teacher as "fair" or even "poor." Obviously the procedure is subjective.

In measuring the work of a teacher, the superintendent may use a measuring instrument or "scale" by which to direct his judgment. For example, his measuring instrument may include the following points:

1. Instructional Skill
2. Discipline
3. Co-operation
4. Community Interest
5. Professional Interest

These five points, while they may be said to constitute a "scale," do not constitute a scale in the same sense that we use the term "scale" in measuring distance. Each point may be called a "unit" in the scale but no one of them is a unit with a definite value as is a foot or an inch.

We may contrast the measurement of quantity with the measurement of quality in the following way:

Measurement of Quantity

1. A zero point
2. Units of definite value
3. Units derived from common experiences
4. Entire quantity measured
5. Objective

Measurement of Quality

1. No zero point
2. Units of indefinite value
3. Units derived from experiences of individual
4. Samples measured
5. Subjective

Measurement in Education both Quantitative and Qualitative

A teacher of English has asked her students to write a composition. In the evening she sits down to mark these compositions. In doing this she perhaps has in mind two general qualities to be evaluated. These two qualities are the mechanical features of the language in the composition and the subject-matter content. In evaluating the mechanical part of a composition she thinks of such things as spelling, punctuation, capitalization, paragraphing, unity, and coherence. She may even "mark" errors in spelling, punctuation, and so on. When she has finished reading the paper she may record the number of errors of a mechanical sort and then decide whether on the basis of the number of errors the composition is to be marked "95" or "80," or some other value. When she thinks of the subject-matter content she doubtless has in mind either the other compositions written by the same class or some general notion which she has gathered from many preceding experiences as to what the content of a composition on this particular subject ought to be; then she marks the quality of the composition according to her best judgment.

In this procedure we note the following: (1) The teacher has no definite "units" in the "scale" which she

uses in evaluating a composition; (2) she may not use a scale at all, except a rather vague general notion which she has in her mind; (3) she may consider certain objective features of the composition such as spelling and punctuation (which features may be called "units" in her "scale"); (4) the scale which she uses has been built up out of her own experiences in marking compositions; (5) the teacher does not presume to measure all of the pupil's ability in composition but gives attention to one sample.

Doubtless most people would agree that in composition marking, described in the preceding paragraph, an effort is made to measure the quality of the composition, but in doing so certain quantitative characteristics are considered. That this is true of measurement in school work may be further illustrated.

A teacher of arithmetic has prepared an examination consisting of fifteen problems. She gives this examination to her class and then proceeds to mark the papers. In marking them she takes into consideration the number of problems correctly solved by each pupil. She gives a higher mark to a pupil who has solved ten problems correctly than to one who has solved only eight problems correctly. She may give no credit for the use of a correct method but may mark entirely on the basis of correct answers—a process of measuring quantity, the quantity measured being the number of problems correctly solved. As a matter of fact, however, this is a method of measuring the ability of a pupil in the phase of arithmetic with which the problems deal. If the teacher should report to us that John's mark was "8" on this examination, it would be meaningless unless we knew something about how many problems there were in the test, how difficult they were, how many problems the other pupils in the class solved, and just how the teacher marked papers. The ex-

amination does not presume to measure *all* of a pupil's ability in arithmetic, nor does it presume to measure *all* of his ability in the particular phase of arithmetic with which the problems deal. In selecting the fifteen problems, the teacher used her judgment as to what types of problems she might best use in getting some indication of her pupils' ability.

We can think of a test in English made up of a very large number of specific questions involving little or no doubt as to the correctness or incorrectness of the answers. It would still be true, however, that there would be both a quantitative and a qualitative element in the process of measurement, the quantitative element entering in connection with the number of questions, and the qualitative element in connection with the judgment of the teacher as to what questions should make up the test. In a test in arithmetic, consideration might be given to neatness, to the choice of methods, and to the clarity with which pupils express their reasoning. Such procedure would reverse the comparison made between English and arithmetic, as presented above.

From the illustrations which have been presented, it seems clear that in the ordinary, informal work of measuring results of instruction we have both quantitative and qualitative elements. The next step is to see how these two elements are present in the measurement of results of instruction by use of the more formal instruments for measuring such results.

The General Character of Formal Tests for Measuring the Results of Instruction

In connection with the last topic we studied the general character of the informal, ordinary test used by teachers

in measuring the results of their instruction. During the last twenty-five years a large number of tests have been developed, of a much more formal character than those prepared from time to time by teachers for use in their own schools.

These more formal tests are, in most cases, made up of a number of "items." The items correspond to the "questions" in the ordinary test and are spoken of as "items" rather than questions because they are not usually expressed in question form. These items are chosen with extreme care. The procedure employed in their choice is usually somewhat as follows. A number of typical textbooks or courses of study, or both, are analyzed and the various elements of subject matter pertinent to the ability which the test is to measure are listed. From this list a number of elements are chosen and each one is represented by an "item." The items are expressed as clearly as possible. The number of items is determined, for the most part, by the time allotted for the test, usually that of an ordinary class-period of from twenty to fifty minutes.

The items are put in a way that makes it possible for the pupil to express his answer with a minimum of writing. The items are expressed in various ways (described in Chapter III) so that the answer of the pupil is correct or incorrect just as definitely as in the spelling of a word or in the answer to a problem in arithmetic.

When the test is in this form it is "tried out" in a number of schools in order to make sure that the items are clearly expressed, that the length of the test is satisfactory, and that the test as a whole is easy enough to enable the poorest pupil to show what he can do, and hard enough to permit the best pupil to show how much he can do in this particular subject or division of the subject.

When the test is satisfactory in all these respects it is

used in a large number of typical schools in order to determine the average scores (the norms) made by pupils in each of the grades in which the test is to be used. Then the author of the test prepares detailed directions for giving the test, for marking the papers, and for interpreting the scores of pupils. The test is then published and put on the market. It is with such tests that this text deals for the most part.

In connection with the use of a measuring instrument such as the test just described we have the following: (1) units (items) of indefinite value; (2) a quantitative basis for judging the accomplishment of the pupil—the pupil's score; (3) no zero point, but instead a standard or norm with which to compare the pupil's score; (4) the selection of items (on a basis that is partly quantitative but also partly qualitative); (5) only a part of the pupil's total ability in the subject measured; (6) norms determined from the average condition.

The following comparison may assist in making clear the vital differences that obtain between measurement in the physical world and measurement in the school so far as both have been developed at the present time.

<i>The Yardstick</i>	<i>A Formal Test</i>
1. A zero point	1. A "standard" or "norm"
2. Units of definite value	2. Units of indefinite value
3. Units derived from common experiences	3. Units derived from common experiences
4. Entire quantity measured	4. Samples measured
5. Objective	5. Objective in large part but subjective to a certain extent

When we analyze the above comparison we see the significance of certain facts. In the case of the yardstick we have a "zero point" as a basis to which to refer all meas-

ures of length. This zero point or value is so common in our experience that we do not need to think of it when we are told, for example, that a board is two feet long. However, when it comes to the formal test the place of this zero point is taken by a "norm." This norm may be "74" for one test, "85" for another, and so on. In connection with one and the same test it may be 74 for one grade, 80 for the next grade, and so on. This means that the score of a pupil on such a test is meaningless until we compare it in a definite way with the norm for his grade.

The "units of definite value" found on the yardstick means that a stick which is 4 feet long possesses twice as much length as one that is 2 feet long. However, a score of 80 on a formal school test in arithmetic does not necessarily indicate twice as much ability in arithmetic as a score of 40. Since the units in this test have no definite value, and usually no known relative value, the 80 score may have almost any ratio to the 40 score. This means that the difference in arithmetical ability represented by scores of 70 and 80 may not be at all the same as that represented by scores of 30 and 40—the difference of 10 in the one case being greater or less than that in the other.

While the units in the case of both the yardstick and of the formal test are derived from common experiences, those in the yardstick are much more generally accepted, perhaps because they are more "natural." This gives us very few measuring instruments for distance, but many for ability in arithmetic, many others in reading, and in all other subjects.

Another thing that gives us a large number of measuring instruments (tests) in education is the fact that ability in any one subject is complex. Ability in reading, for example, is made up of (a) the ability to recognize words, (b) the ability to know the meanings of words, (c) the

ability to understand the thought relationships among sentences, (d) the ability to organize for definite purposes what is read, and perhaps other abilities, the total number depending upon the point of view from which the analysis is made. The measurement of reading ability *in toto* may be compared to measuring a block of wood as to its (a) length, (b) breadth, (c) area, (d) volume, and (e) density, for each of which measurements we would use a different measuring system.

Since the entire quantity is measured in the case of distance and since the measurement is objective, the total error in such measurement is slight. On the other hand in the case of a formal test in arithmetic, since only a sample of a pupil's ability is measured and since the measurement is in part subjective, the possibility of error is large. This error is so unavoidable and is so large that in connection with most tests of this sort the "probable error of the score" is given and any individual score is interpreted in the light of this probable error.

Such then is the present status of measurement in the field of education. The student who is to develop a sane and intelligent attitude toward educational tests should understand the fundamental weaknesses as well as the elements of strength in the system. He should develop a wholesomely critical, constructive attitude, rejoicing over the fact that he has an opportunity to work in a great field of human endeavor while this field is in its pioneer stages. Measurement in education will probably never, because of the intrinsic nature of the material to be measured, become as accurate and as satisfactory as is the measurement of things in the physical world, but it will surely be perfected to a far greater degree than is the case at present. If this is ever achieved it will be accomplished largely through the criticisms and suggestions of the thou-

sands of classroom teachers who use tests, as well as through the efforts of scientific workers in the special field of measurement.

The student should become acquainted with specific tests, with the various types of tests, with the ways in which tests are constructed, with the various criteria by which tests are selected, with the full theoretical and practical significance of such terms as validity, objectivity, and reliability, and with the various statistical procedures that are employed in the practical use of tests. The purpose of this book is to help the student to make this acquaintance.

The Beginning and the Development of the Formal-Test Movement

The test movement proper has a very brief history. In the 1890's Dr. J. M. Rice, a practicing physician, became interested in the work of the schools in connection with its effects on pupils and teachers. His idea was that many teachers and pupils suffered more or less from nervous disorders brought on by the fact that, apparently, perfection was expected of them—perfection in spelling, in arithmetic, in geography. Dr. Rice thought that since perfection obviously could not be attained, it would be advantageous to have some sort of reasonable standards set up which could be attained. Accordingly, he proceeded to devise some tests in spelling and in certain other subjects and to give them in a number of schools in order to ascertain just what could reasonably be expected of pupils.

Dr. Rice's work met with a storm of opposition largely because he deduced from his findings certain conclusions that were very obviously unwarranted. However, it was apparent that his ideas and his plan in general had pos-

sibilities. The work was taken up by other men and gradually the fundamental features of Dr. Rice's procedure were developed and applied more carefully and more scientifically than was the case in his pioneer work. To Dr. Rice, however, more than to anyone else, belongs the credit for launching the movement.

At first and for many years, tests of a formal character were limited to those subjects or phases of subjects that are of a mechanical nature—spelling, the fundamental processes in arithmetic, place geography, etc. Gradually, however, the work was extended into the less mechanical subjects, until now we have formal tests in all subjects and for many characteristics of pupils, such as general intelligence, special aptitudes, moral judgment, personality traits.

In the early days of the formal-test movement the use of such tests was largely a matter of curiosity—curiosity as to how the results compared in achievement with the “standard” score. Now the use of tests has become a component part of school procedure. Many school systems have set up definite testing programs. They use tests for various purposes such as the following: (a) to keep the achievement of the school as a whole on as high level as possible; (b) to measure the general results of different procedures; (c) to check up on the classification of pupils; (d) to determine the particular difficulties of individual pupils; (e) to compare the achievement of the school in the various subjects; (f) to measure the general ability of pupils and to compare their accomplishment with their general ability. Many cities and many states employ specialists in the work of testing.

PROBLEMS

1. What different types of measurement would you classify as "educational measurement"?
2. Before a unit of work in arithmetic was studied, a teacher gave a test of 100 points. Charles made a score of 40 and Mary a score of 62. After the unit was completed the same test was given again and Charles scored 69 and Mary 87. Can you tell which has made the greater gain?
3. Do tests have any other function than that of improving the work of the school?
4. Can you think of a field other than education in which measurement has been made from an arbitrary zero point?
5. Is it strictly true that in quantitative measurement the entire quantity is measured, or can you think of exceptions?

BIBLIOGRAPHY

- Caldwell, O. W., and Curtis, S. A., *Then and Now in Education*. World Book Company, 1925.
- Hawkes, H. E., Lindquist, E. F., and Mann, C. R., *The Construction and Use of Achievement Examinations*. Chapter I. Houghton Mifflin Company, 1936.
- Lang, A. R., *Modern Methods in Written Examinations*. Houghton Mifflin Company, 1930.
- Odell, C. W., *Traditional Examinations and New-Type Tests*. Chapter I. D. Appleton-Century Company, 1928.
- Orleans, Jacob S., *Measurement in Education*. Chapter I. Thomas Nelson and Sons, 1937.
- Tiegs, E. W., *Tests and Measurements for Teachers*. Chapter I. Houghton Mifflin Company, 1931.

CLASSIFICATION OF TESTS

1

1

Chapter II

CLASSIFICATION OF TESTS

CONSIDERABLE CONFUSION has at times arisen in the minds of inexperienced test users concerning the functions which tests are expected to perform, and particularly concerning the functions of the various types of tests. One of the errors of inexperienced users has been in regarding a test as a remedy for situations of one sort or another. Remedial measures are frequently suggested by test results, but the tests themselves never correct conditions. Tests accomplish their objective when existing conditions have been revealed, but the teacher must bear in mind that when the tests have been administered and scored, the work of improving conditions can be begun and is in no sense completed. This does not mean that tests have little value. On the contrary, the analysis of existing conditions is as valuable to the teacher or superintendent as is the diagnosis of ailment to the physician. The diagnosis does not cure, but proper remedies can be applied only after the nature of the disease is disclosed.

Tests are used to reveal the existing state of affairs in many situations. For example, a superintendent may wish to know how his school system compares with other school systems of similar size in general scholastic achievement or in some particular school subject. Again, he may wish to know which schools in his city are doing the most

effective work. A principal may desire information concerning the mental ability of a third grade group which is having considerable difficulty in learning to read fluently. Or a teacher may be eager to ascertain just what particular number combinations remain to be automatized. All of these questions come within the legitimate demands which may be made on a testing program, provided the testing program is so organized as to point to the answer to such questions.

A. TESTS CLASSIFIED AS TO FUNCTION

Most of the tests so far devised, with the possible exception of certain temperament and personality tests, may be classified as aptitude tests—sometimes called prognostic tests, or as achievement tests.

1. Aptitude Tests

General Intelligence Tests. The most commonly used aptitude tests are the so-called tests of intelligence or of mental ability, the chief purpose of which is to determine general aptitude or native capacity. As is pointed out later (in Chapter XI) these tests do not measure all of one's capacity, but in the main one's capacity for academic work. In general, the items for tests of this sort are not chosen because they represent abilities which children should acquire in school, but because they represent abilities which the normal individual will possess, despite lack of normal opportunities, simply because the human being is so constituted that, at least as a child, he must learn something almost regardless of his environment. For example, most normal children of six years will be able to state with reasonable accuracy the difference between (a)

a bird and a dog; (b) a slipper and a boat; and (c) wood and glass (as they are asked to do in the Terman-Merrill Revision of the Stanford-Binet Scale) without having been taught specifically.

In other words, there are certain bits of information and certain abilities which the average child acquires because he lives in the United States, or which he possesses because he is a member of the human race. Such information and such abilities constitute the logical materials for intelligence test items. The following items from the Henmon-Nelson Tests of Mental Ability were found to distinguish well between children of superior and inferior intelligence, as the term is ordinarily used.

Which word does not belong with the others?

(1) pen, (2) rock, (3) paper, (4) ink, (5) pencil

At a dinner there is always:

(1) a tablecloth, (2) cutlery, (3) company, (4) merriment, (5) food

The high were rugged and bare. A word for the blank is:

(1) valleys, (2) seas, (3) mountains, (4) lights, (5) storms

Which word does not belong with the others?

(1) Monday, (2) April, (3) January, (4) May, (5) June

Which of these words comes first in the dictionary?

(1) apple, (2) long, (3) winter, (4) snow, (5) peach

While such items do appear to measure accomplishment rather than native capacity, there is as yet no known method of measuring intelligence without using accomplishment as an intermediary.

Tests of Special Aptitude. Almost everyone is acquainted with some individual who, while not showing particular promise in all lines, is exceptionally able in a certain type of work. Sometimes the proficiency may lie

in a given school subject such as art, mathematics, or music; sometimes unusual proficiency is shown in mechanics or some other field not ordinarily pursued in the public schools. All of us are familiar also with instances in which a person who has not been successful in a given field has, by changing his line of endeavor, become much more successful and happy in his work. Workers in the field of aptitude testing have seen the advantage in discovering the individual's talents as early as possible, so that his school work may be pleasant and profitable and his choice of a vocation such as to make him a happy and useful citizen. Special aptitude tests are not designed to measure pupils' achievements but to predict probable success in a given school subject or in a given vocation. The Stenquist Mechanical Aptitude Tests¹ and the Seashore Musical Talent Tests² are good illustrations of such tests which help teachers or parents in guiding children into the proper courses and vocational choices. In some instances also prognosis tests for a given school subject have been devised, utilizing in the main the achievement in an allied field. Still other tests attempt to predict what the pupil will be able to do by giving him a small sample of the work required. This plan is followed, for example, in the Orleans-Solomon Prognosis Test,³ where the pupil is tested on a miniature Latin learning situation.

2. *Achievement Tests*

While most aptitude tests are designed to indicate what the pupil will be able to do because of his inherited capacities and his interests, the function of achievement tests is

¹ Published by World Book Company.

² Published by The Columbia Graphophone Company.

³ Published by World Book Company.

to determine what the pupil has learned as a result of a period of training. While most writers classify such tests into two groups; namely, survey tests and diagnostic tests, Brueckner and Melby⁴ have made five divisions which appeal to the writer as being serviceable and more indicative of their function. The five groups are: (1) general survey tests, (2) subject achievement tests, (3) analytical subject tests, (4) diagnostic tests, and (5) curriculum tests.

General Survey Tests. As the name implies, tests of this character are designed to measure the general achievement of pupils. This may be done by measuring separately the achievement in various school subjects and then combining the scores, as is done in the case of the Stanford Achievement Tests⁵ and other so-called "batteries" of tests, or by arranging the items of all subjects in the "omnibus cycle" plan as is done in the Otis Classification Test.⁶ Under the latter arrangement all of the easiest items, whether they be in spelling, language, arithmetic, geography, or one of the other subjects, appear first, with the items gradually increasing in difficulty. The Otis test also contains an intelligence test for those who wish to use it.

Tests of this type point out the general state of scholastic progress in the school system where they are employed. They are particularly helpful to administrative officials, indicating as they do how the school system compares in general achievement with other school systems, or how the different schools within a given system compare with one another. In most of these tests the knowledge or skill required in a given subject is not tested in sufficient detail

⁴ Brueckner and Melby, *Diagnostic and Remedial Teaching*, p. 69.

⁵ Published by World Book Company.

⁶ Published by World Book Company.

to give the teacher an adequate index of the individual pupil's achievement.

Subject Achievement Tests. Because they confine their attention to a narrower field; namely, a given subject, tests of this type generally give more reliable information concerning the individual pupil's status in a given subject. As an illustration of tests of this type one may mention the Denny-Nelson Tests in American History⁷ and the Woody-McCall Mixed Fundamentals in Arithmetic Scales.⁸ Their chief usefulness lies in noting how the achievement of individual pupils, as well as the class as a whole, compares with the achievement of pupils of the same status in other schools, as indicated by the published norms. The tests are not so designed that particular weaknesses are apparent, however. If, for example, a pupil arrives at the wrong answer to a problem in arithmetic, no provision is made for finding the exact error or the source of it.

Analytical Subject Tests. Some of the subject achievement tests are so constructed that a general analysis of pupil strength or weakness may be made. Illustrative of tests of this type is the Wiedefeld-Walther Geography Test⁹ which is so divided that a teacher can tell, in a general way, by noting how a pupil's score varies from the published norms, whether the pupil has greatest difficulty with reading geographic materials, with organization of materials, with map and graph reading, with geography vocabulary, with knowledge of geographical relationships, or with knowledge of place geography. Tests of this type have been developed most frequently in reading, where exact diagnosis is very difficult, especially in groups, but

⁷ Published by World Book Company.

⁸ Published by Teachers College, Columbia University.

⁹ Published by World Book Company.

where the results of such analytical tests prove helpful in determining the type of reading which pupils need to emphasize. Reading tests of this type are the Gates Silent Reading Tests,¹⁰ the Nelson Silent Reading Test,¹¹ and the Sangren-Woody Reading Tests.⁹ Tests of this sort are also available in history and grammar.

Diagnostic Tests. Many of the so-called diagnostic tests are in reality either analytical tests or simply survey tests, since they fail to make adequate provision for the determination of exact weaknesses or deficiencies. To be truly diagnostic a test should point out the exact pupil weakness which needs to be corrected. While this is comparatively easy in the case of certain arithmetical fundamentals, it is exceedingly difficult in many other phases of arithmetic and in most other subjects. Even where a given specific difficulty is found, as in the case of the pupil who persists in securing seven as the answer to the combination $5 + 3$, the cause of the difficulty is not always apparent and hence diagnosis is not complete.

Many tests may be used for survey purposes, for analysis, or for diagnosis. A good example of this is afforded by the Pribble-McCrory Diagnostic Tests in Practical English Grammar.¹² The user of this test may secure the total score, and by comparing this with the published norms, determine whether his pupils are doing relatively well or rather poorly as compared with the pupils of other schools. He may analyze each paper further and discover that certain pupils have difficulties with verbs, others with pronouns, and still others with nouns. By noting which pupils have these various difficulties, the class may be broken up into small groups for drill purposes. But the

¹⁰ Published by Teachers College Bureau of Publications.

¹¹ Published by Houghton-Mifflin Company.

¹² Published by Lyons and Carnahan.

most valuable information will be obtained by the teacher who notes each item missed by each pupil and plans remedial instruction for each pupil in such a way that his particular difficulty will be overcome. The individual item study is a diagnosis, or at least a beginning of the diagnostic process.

Curriculum Tests. Tests bearing this name are designed to be used at the completion of a given unit of the curriculum in a particular subject. Because of the fact that the curriculum of one school system often varies to a considerable extent from the curricula of other systems, these tests are not very numerous as yet. When there is more standardization of the materials of instruction, we may look for their development, for it is stimulating for both pupils and teacher to be able to test achievement at periods of, say, every four or six weeks, instead of only once or twice each year. As it is, most curriculum tests have been produced for a given city only, although a few standardized curriculum tests are available in arithmetic and spelling particularly, since curricula vary somewhat less in these subjects. The motivating value of repeated testings has been found to be so great by Henmon¹⁸ and other investigators, that wider use of such tests would no doubt be very profitable. One of the best known of the curriculum tests is the series devised by Brueckner¹⁴ for arithmetic processes. In this series there is a set of ten tests for each of the grades 3 to 8 inclusive, which have been standardized for each month. Since curricula vary somewhat even in arithmetic, the teacher who uses the tests may find it necessary to give them at irregular inter-

¹⁸ Henmon, V. A. C., Improvement in School Subjects Throughout the School Year, *Journal of Educational Research* 1:81-95, February 1920.

¹⁴ Brueckner, *Curriculum Tests in Arithmetic Processes*, John C. Winston Company.

vals rather than at the end of each month as intended, but their use even in this way can be recommended.

B. TESTS CLASSIFIED AS TO FORM

The Essay Test

Before 1915 the essay type of examination was used almost to the exclusion of other types. Since that time other types of tests have come to be used in increasingly large number but the essay examination is still widely used. The following set of questions in American history, which were designed for a sixth grade class, will serve as an illustration of the kind of test under consideration.

AMERICAN HISTORY

(Answer any five questions)

1. Describe briefly the early history of the following colonies: Rhode Island, Delaware, Massachusetts, and New York.
2. Name five weaknesses of the Articles of Confederation.
3. Describe how we acquired each of the following: Louisiana Purchase, Texas, Florida.
4. Discuss briefly Magellan's trip around the world.
5. Discuss the causes of the Revolutionary War.
6. Discuss the meaning of the Monroe Doctrine.

In taking this examination one pupil wrote the following set of answers.

AMERICAN HISTORY

1. Omitted.
2. Five weaknesses of the Articles of Confederation:
 - (1) Too brief
 - (2) Applied mainly to common people and exempted the main part of the higher class

- (3) Enforcing the laws wasn't adequately done
 - (4) The Federal class was more in power
 - (5) The Government, under the Articles, wasn't democratic enough
3. We bought Louisiana from Napoleon in 1803, during Jefferson's administration.
- We annexed Texas in 1848, after she had set up a government of her own attempting to free herself from Mexico. The war with Mexico followed as a result of this.
- We purchased Florida from Spain about the time of the Spanish-American War.
4. Magellan made the first trip around the world shortly after Columbus' voyages. He left Spain sailing west and returned from the east, sailing around Africa. He accomplished what Columbus failed to do.
5. The causes of the Revolutionary War:
- The chief cause for war was the desire of the colonies for independence. There were many other instances that led up to the immediate action, as: Stamp Act, Intolerable Acts, Boston Massacre, Boston Tea Party, until the actual firing on Fort Sumter.
- England wished to keep the colonies as she had taken them from the hands of other European countries and wished to use them as a trading center and a place to sell her goods.
6. The Monroe Doctrine was a doctrine formed for the sake of keeping foreign countries, especially Latin America, out of the colonies' business, and they agreed to keep out of the foreign countries' business. It was later changed to the effect that it applied to all countries as well as Latin America.

In order to determine the amount of agreement among teachers in grading papers of this sort, the above paper was mimeographed and submitted, with the questions, to

forty-three summer school students who had been teaching American history in the sixth grade during the previous year. The total scores assigned by these students are indicated in the following table.

90-99	1
80-89	5
70-79	5
60-69	11
50-59	12
40-49	7
30-39	2

The exact range in grades assigned was from 38 to 96. In other words, a pupil under the instruction of one of these teachers would have passed the examination with flying colors; under another teacher, he would have failed the same examination miserably. Neither is there any very good method of determining which of the graders, if any, had done an accurate piece of work. Many more extensive studies of this type have been made, the pioneer work having been done by Starch and Elliot, using papers in English, geometry, and history, but all of them lead to the same general conclusion; namely, that teachers vary widely with respect to the grades they assign to the same paper.

If we examine the grades assigned to a single answer, as number 6, in the above paper, we find them varying as indicated in the following table.

<i>Grade Assigned</i>	<i>Number of Graders</i>
20	9
19	0
18	4
17	3
16	1
15	5

<i>Grade Assigned</i>	<i>Number of Graders</i>
14	0
13	1
12	0
11	0
10	9
9	0
8	0
7	0
6	0
5	4
4	0
3	0
2	0
1	0
0	7

Other studies indicate that teachers not only disagree among themselves, but that when the same teacher grades a given paper after an interval of time, and without a record of her earlier grading, the second is likely to differ considerably from the first. One criticism of the essay test is, then, that it is lacking in objectivity, that is, in that quality which enables two or more graders working independently to assign the same grade.

Affords Limited Sampling. In the typical essay test the pupil is usually required to answer five or ten questions. In answering these he frequently spends one or more hours. The examiner is thus able to sample only a very limited amount of the pupil's knowledge. It must be remembered that in practically all testing situations only a *sampling* of the pupil's knowledge or ability or skill is afforded, and from this sample the general ability of the pupil is inferred. With one hundred or more such samples, which could be obtained in the same amount of time if objective test items were used, one would secure a much more adequate sampling.

It should be pointed out, however, that a different type of sampling is involved in the essay examination. We may use the term *intensive* to apply to the sampling afforded by the essay test, since the teacher is able to determine in some detail the pupil's knowledge of the situation involved. The sampling obtained by the objective test is more *extensive*, that is, covers more samples but goes less into detail in each case.

Has Low Reliability. Failure to secure an adequate sampling is one of the reasons that most essay examinations have low reliability. By the reliability of a test is meant its consistency of measurement; that is, a test is considered to be reliable if it measures consistently whatever it sets out to measure. However, methods of determining reliability will be discussed in a later chapter. The lack of agreement shown by teachers in grading essay type papers is an indication that the consistency of measurement of such a test cannot be high. Moreover, irrelevant factors frequently play too large a part in the determination of a pupil's grade. If, for example, the time limit for the test is short, the pupil who writes slowly is at a disadvantage. Teachers who grade such papers are often influenced by the quality of handwriting, by their own feelings of fatigue or well-being, and sometimes by atmospheric conditions. Obviously, grades which are influenced by so many extraneous factors are not likely to be consistent.

The Objective Test

Although forty or more different types of objective test items have been devised ¹⁵ and more are being added from time to time, only a few of them have come into common

¹⁵ Hawkes, Lindquist, and Mann, *Achievement Examinations*, Houghton Mifflin Company, 1936, p. 107.

use. Those mentioned below are those most frequently found in standard tests now available.

True-False Statements. So frequently has this type of item been used that many persons consider the terms "objective tests" and "true-false tests" to be practically synonymous. This type of item, in which the pupil is asked to indicate whether the statement is true or false, has been so greatly abused by inexperienced test-makers that many persons have been turned against the entire objective testing movement. The following items will serve as illustrations of the type of item under consideration.

Plymouth Colony was established in 1607.	True	False
--	------	-------

Jackson was the first president to make excessive use of the Spoils System.	True	False
---	------	-------

This type of test has certain merits which may be listed as follows:

1. It is relatively easy to construct.
2. It is entirely objective.
3. It permits of easy and rapid scoring.
4. It permits of wide sampling because very little time is required for each item.
5. It can be used in most subjects.
6. It can be so constructed as to test for ability to reason as well as for facts memorized.

The chief objections which have been raised to this type are:

1. That it is open to guessing and to chance effects to a very considerable extent;
2. That in subjects where much of the material is controversial, it does not serve very well;
3. That it is not so easy to construct as is commonly thought if care is exercised in the avoidance of ambiguous or partly true and partly false items;

4. That the false statements may leave, in the mind of the immature pupil especially, an erroneous impression.

Yes-No Questions. To meet the fourth objection to the True-False statement; namely, that false items may, because of their positive character, leave an erroneous impression with the pupil, the Yes-No question has been rather commonly employed. To illustrate, the True-False statements appearing above might be changed to read:

Was Plymouth Colony established in 1607? Yes No

Was Jackson the first president to make
excessive use of the Spoils System? Yes No

Studies which have been made do not reveal that misinformation is acquired from the True-False test. However, it may be that questions put in this form are approached with a different psychological attitude from that in which a True-False statement is studied. Whether one presents a more typical problem situation than the other is an open question. Except that they may be less likely to leave pupils with erroneous impressions, the Yes-No questions have all of the advantages and limitations of the True-False items.

Multiple Choice—Single Response. As an illustration of this type of item we may change one of the illustrations used above to read:

Plymouth Colony was established in:

(1) 1607, (2) 1620, (3) 1636, (4) 1700, (5) 1755.

Such items have the advantage of being:

1. Fairly easy to construct;
2. Entirely objective;
3. Rather highly reliable;
4. Less subject to guessing and chance effects than the True-False or Yes-No types.

On the other hand, such items are:

1. Likely to consume considerable space and to involve a considerable amount of reading on the part of the pupil;
2. Difficult to construct so that the incorrect responses are not too obviously wrong;
3. Likely to become purely factual.

This last objection is not altogether valid since these items can be so constructed as to test reasoning as well as facts. The danger is, however, that when this is done, the questions become excessively long.

Multiple Choice—Plural Response. In the illustration above the pupil was to select only one answer since only one could be correct. In the item below, the pupil must designate two answers.

Two cities in the following list are capitals of states. They are: (1) Milwaukee, (2) Albany, (3) Los Angeles, (4) Des Moines, (5) Chicago

Any number of correct responses may be included, but it has been the common practice to indicate to pupils the number of correct answers which are to be found. Except that the opportunity for guessing and for chance may enter in, to a somewhat greater degree, unless more responses are provided, such an item has the merits and limitations of the single response form.

Recall-Completion Tests. As an illustration of this type our original examples may be modified so as to read:

Plymouth Colony was established in

The first president to make excessive use of the Spoils System was

This form differs from those just discussed in that the correct answer or answers are not supplied. Psychologically, then, the pupil is confronted with a situation de-

manding recall rather than recognition. Among the advantages one may mention the following.

1. It is relatively easy to construct.
2. It is almost entirely objective.
3. Guessing and chance factors are almost eliminated.
4. The form of question is more nearly akin to that employed in classroom recitations.

On the other hand it has been criticized as being:

1. Too factual in character;
2. Not entirely objective;
3. Somewhat difficult to score;
4. Somewhat difficult to construct if one is to approach perfect objectivity;
5. Subject to the speed of pupils' writing.

Matching Exercises. This type of exercise has characteristics which make it differ from most tests. The following exercise will serve as an illustration.

Below you will find listed a number of inventors and a number of inventions. Write in the parenthesis at the right the number of the person to whom the invention named is usually credited.

<i>Inventor</i>	<i>Invention</i>	
1. Samuel F. B. Morse	The spinning jenny	()
2. Edmund Cartwright	The phonograph	()
3. Elias Howe	The steamboat	()
4. Eli Whitney	The telegraph	()
5. Robert Fulton	The telephone	()
6. John Ericsson	The power loom	()
7. S. M. Babcock	The reaper	()
8. James Hargreaves	The cotton gin	()
9. Alexander Graham Bell	The steam engine	()
	The sewing machine	()
10. Thomas Edison		
11. James Watt		
12. Cyrus McCormick		

Among their advantages it is well to note that:

1. They can be used to measure judgment or mastery of fact;
2. Chance and guessing factors do not enter in if ten or more pairs or incomplete matchings are used;
3. They are easily scored.

Among the criticisms voiced against them, those most commonly heard are that:

1. They cannot be used in all subject-matter fields;
2. Exercises containing many pairs are wasteful of pupils' time;
3. Exercises containing few pairs permit chance to enter.

Irrelevant Terms. Test items of this type are not so commonly used in achievement tests as in tests of intelligence. They are particularly useful in testing a pupil's ability to classify persons, objects, or places. The following illustration from an intelligence test will serve.

Which of the following words does not belong with the others? (1) solarium, (2) academy, (3) seminary, (4) college, (5) university

Tests of this sort have about the same merits and limitations as were listed for the Multiple Choice. Like these tests also, they may be so constructed that the pupil is required to select two or more irrelevant terms.

PROBLEMS

1. What factors, other than the correctness of the response, are likely to influence teachers in their grading of essay tests? Is a teacher justified in taking any of these factors into account? If so, which ones?
2. Do you agree with the statement that Yes-No items have all of the advantages of the True-False items? Would, for ex-

ample, the situation in the True-False test more nearly duplicate the situation one encounters in literature designed for propaganda?

3. What is meant by saying that a test is thoroughly objective?
4. As a classroom teacher, which of the types of tests discussed in this chapter would be most helpful to you? Why?
5. If you were the assistant superintendent in charge of elementary education in a large city, what sort of a testing program would you probably like to have in your schools?

BIBLIOGRAPHY

- Brueckner, L. J., and Melby, E. O., *Diagnostic and Remedial Testing*, Houghton Mifflin Company, 1931.
- Greene, H. A., and Jorgensen, A. N., *The Uses and Interpretation of Elementary School Tests*, Longmans, Green and Company, 1936.
- Henmon, V. A. C., Improvement in School Subjects Throughout the School Year, *Journal of Educational Research* 1:81-95, February 1920.
- Ruch, G. M., *The Objective or New-Type Examination*, Scott, Foresman and Company, 1929.
- Smith, H. L. and Wright, W. W., *Tests and Measurements*, Silver Burdett Company, 1928.
- Tiegs, E. W., *Tests and Measurements for Teachers*, Houghton Mifflin Company, 1931.
- .

CONSTRUCTING AND SCORING TESTS

Chapter III

CONSTRUCTING AND SCORING TESTS

IT WOULD BE WHOLLY IMPOSSIBLE to overemphasize the need for care in constructing examinations, regardless of the type one plans to use. Every test worthy of the name should be as adequate a measuring instrument as it is possible to devise. When one considers the care that has been exercised in constructing measuring devices used in the physical realm as compared with the time used by most teachers in constructing their tests, one of the reasons for the inadequacy of educational measurement becomes apparent. It is not suggested that care in construction accounts for all of the differences in these two types of measurement, but it does appear to be an important consideration. Unfortunately the number of times that an educational test can be used, in comparison with an electric meter or a yardstick, indicates that less time can be spent on test construction. It does not indicate, however, that educational measurement is less important or that poor tests can, therefore, be condoned. Every test should accordingly be as good as it can be made, considering the time available for its preparation, the testing time available, and the time and effort that can be expended by the pupils.

The Essay Examination

From the discussion of the essay examination in the previous chapter the reader may infer that the writer is opposed to its use in any situation. As a matter of fact, some persons consider the essay test as a wholly useless device for educational measurement. Such statements are undoubtedly too sweeping. In the first place, the essay examination is familiar to teachers and to most pupils. That there is some advantage in working with devices with which one is familiar can be illustrated from the much more accurate results obtained by the trained surveyor as compared with the novice, though each may use the same set of instruments.

A second advantage of the essay examination is the fact that it is almost the only test that, up to this time, has proved satisfactory for the measurement of ability to organize materials. In many fields, on the other hand, this ability is one of the major objectives of course instruction. The writer is convinced that ability to organize material can be tested more objectively than is done by means of essay tests; as a matter of fact, some fairly successful attempts have already been made in testing this ability objectively. Most of the techniques are, however, quite involved and too cumbersome for the average classroom teacher to use in constructing her own tests. Advocates of the objective tests frequently reply that the mental state of a pupil taking an examination is such that he cannot do himself justice when he attempts to express himself coherently and that what is being measured is really composition ability, which cannot be rated objectively. Some argue, and perhaps with justice, that if practice in organizing material is wanted, it is better to give this practice apart from a testing situation.

Most of the criticisms of the essay test center about the subjectivity; that is, the lack of objectivity of such examinations. Yet, if care is taken in the construction of such tests and in the grading of them, it is possible to decrease the subjectivity to a marked degree. In the history questions listed in the previous chapter (page 45), question number 3 is usually scored much more consistently than is question number 5. By using a uniform set of scoring rules much of the subjectivity may also be eliminated. A study by Kelley leads to the following conclusion: "If we take the position that disagreements of no more than five points are not very serious, almost ninety-five per cent of the 219 pupils were marked with reasonable accuracy when rules were employed, while in the absence of rules but sixty-two per cent showed differences of five points or less."¹

Ashbaugh had experienced teachers grade an arithmetic paper on three occasions without scoring directions, and finally with detailed directions for scoring. When scoring rules were used the total range of scores was only eighteen points (68-86) as compared with fifty-one points on the first scoring, fifty on the second, and thirty-nine on the third. The number of cases varying widely from the median was also materially reduced.²

The teacher who wishes to improve her essay examinations will be guided by having in mind the criticisms that have most commonly been leveled against this type of test. The two most common of these are (1) that the test affords a limited sampling and (2) that scoring is likely to be subjective. The first criticism suggests that a better

¹ Quoted from Ruch, *The Objective or New-Type Examination*, Scott, Foresman and Company, 1929, p. 102.

² Ashbaugh, E. J., Reducing the Variability in Teachers' Marks, *Journal of Educational Research* 9:185-198, 1924.

sampling should be provided and this can be done by the simple expedient of providing for answers which will be shorter and less time-consuming for the pupil to write. If, for example, twenty questions are asked which can be answered in the same length of time as has been required for answering five, a more adequate sampling is likely to result, with due regard to the significance of each item. Likewise, the teacher who will bear in mind, as she constructs her questions, what are the possible correct responses, will be in a position to construct better tests and to score them more objectively than the teacher who neglects to consider this matter until the papers come to her for scoring. The teacher must also be on guard against taking into account irrelevant factors, such as handwriting. In order to offset the danger that papers graded first will receive more favorable consideration because of the teacher's comparative freedom from fatigue, some teachers find it best to grade the first item on all papers, then the second item on all papers, etc.

It is interesting to note that when educators sensed the limitations of the essay test they turned eagerly to the objective examination in the hope of solving their measurement problems. It is possible that, if as much research had been done in improving the essay examination, even more progress might have been made in evaluating the results of teaching. As it is, we have many more worthwhile suggestions for the improvement of objective tests than we have for the essay examination.

True-False Statements

This type of objective test item has probably been more abused than any other and for the reason that the ease with which it can be constructed is so deceptive. Many

teachers have followed the practice of reading through the textbook material to be covered and selecting, more or less at random, sentences appearing in the copy. By modifying a word or two, or by inserting a negative, they are able to make the needed number of false statements with a minimum amount of effort. In following this procedure they are frequently unaware that a statement which is removed from its context may have a quite different meaning or, at least, be capable of a quite different interpretation from that which was intended. This method proves advantageous to the pupil who is adept at photographic recall of the printed page and sometimes penalizes the student who has gained a more comprehensive understanding of the subject matter without minute attention to the author's mode of presentation. It is also likely that some of the statements thus selected will be concerned with minor details which are not particularly pertinent to the objectives of the course. A much better technique is for the teacher to set down in considerable detail the objectives of instruction for the unit under consideration and then formulate in her own words the statements which cover these objectives. The following suggestions will probably be of value in constructing items of this type.

1. Make provision for recording responses which will be as convenient as possible for both scorer and pupil. There are several different ways of directing pupils to indicate their responses. Some teachers prefer to have pupils write the words *True* or *False*. This involves more writing, and hence more time than is needed, for the pupil, and probably slows up the scorer. Others use the letters *T* and *F*, but these letters are so similar in construction as to cause confusion in many instances. Some use the plus sign for the true statements and the minus sign for false statements.

This system is fairly satisfactory but confusion sometimes results even here when a pupil wishes to change a plus to a minus and does so by simply emphasizing the horizontal marking. The techniques indicated under A and B below have been found by the writer to be quite satisfactory. Allowance must be made, however, for the maturity of the pupils.

A. Directions: In the blank before each statement place a + sign if the statement is true and a 0 if the statement is false. *Do not guess.* Items 1 and 2 are correctly marked.

- . 0 . 1. Maine entered the union as a slave state.
- . + . 2. The Constitution provided for a government of three branches.
- 3. President Taylor assisted in establishing United States banks.

B. Directions: Encircle the T before all true statements and the F before all false statements. *Do not guess.* The first item has been marked correctly.

Ⓓ F 1. John Adams was the second President of the United States.

T F 2. Harrison served only one month as President.

2. Avoid use of the language of the textbook. This is especially important if removing the statement from its context makes it ambiguous. As was pointed out above, the use of textbook language encourages rote learning as contrasted with ideational learning.

3. See that the number of true statements is about the same as the number of false statements. To use a great preponderance of either true or false statements is likely to decrease the reliability of scores corrected for chance. On the other hand, if the teacher always uses an *identical* number of each type, bright pupils may resort to counting to increase their chances in guessing correctly.

4. Use good English. This would seem to be an un-

necessary admonition, yet it is surprising to find how often this precept is violated.

5. Avoid long statements, particularly compound sentences.

6. Avoid negatives, particularly more than one in the same statement.

Bad: The small states were not less eager to adopt the Constitution than were the large states.

Better: The small states were more eager to adopt the Constitution than were the large states.

7. See that the vocabulary is adjusted to the level of the pupils.

8. Avoid "trick" or "catch" questions.

9. Avoid statements that are ambiguous or partly true and partly false.

10. Avoid having in the same test two items, one of which suggests the answer to the other.

11. Be sure to emphasize, by position or otherwise, the crucial element in a statement.

Bad: The Articles of Confederation, adopted in 1789, became the first rules of government for all of the states.

Better: The Articles of Confederation, which were the first rules of government for all of the states, were adopted in 1789.

It is assumed in the above illustrations that the crucial element is the time of adoption of the Articles of Confederation. If some other phase is the crucial element, its position should emphasize it.

12. Avoid the use of such words as "all," "none," or "always" more frequently in false than in true statements. There is no objection to the use of such words if they can be used as frequently in true as in false statements. This is usually difficult since there are relatively few rules which do not have some exceptions. Pupils soon learn that such words are commonly associated with falseness and hence guess more of them correctly than pure chance would permit them to do.

13. Avoid ambiguous terms such as "large," "small," "important," etc. Where comparisons are to be made, see that the comparisons are direct or stated in quantitative terms. In the previous chapter we used as an illustration of the True-False item the following statement:

Jackson was the first President to make excessive use of the Spoils System.

The use of the word "excessive" leaves the merits of this item open to question since what is excessive is subject to various interpretations.

14. In scoring, indicate the pupil's score as the difference between the number right and the number wrong. The usual scoring formula for two-response items is $S = R - W$, or the score equals the number right minus the number wrong. Omitted items are not counted as wrong.

15. In general, make each statement independent of other statements in the test. This rule may sometimes be violated without injury to the test but in general it is a safe rule to follow since, unless care is exercised, such items will be confusing to pupils.

The above suggestions will be found helpful in the construction of Yes-No items and, in general, in making all types of two-response tests

Multiple Choice Items

This type of item is one of the most useful yet devised. It has grown in popularity among test makers because of its adaptability to various types of subject matter and its possibilities for testing reasoning as well as factual information. There is also the advantage that the chance factor is considerably less than that present in the True-False test. It can be made completely objective in scoring and can be scored very quickly. It does, however, require con-

siderable care and time in preparation. The following suggestions may be helpful.

1. Arrange the test in such a way as to be convenient for both the scorer and the pupil. The following directions and form are in common use and are probably as satisfactory as any.

Directions: In the parentheses at the right of the page write the number of the correct response. There is only one correct response for each item.

1. The presiding officer at the Constitutional Convention was: (1) Adams, (2) Washington, (3) Patrick Henry, (4) Franklin, (5) Jefferson. (2)

Some test makers prefer to place the parentheses before the item, as in the following:

- (4) 1. Brigham Young and the Mormons established themselves permanently in: (1) New Orleans, (2) Detroit, (3) Denver, (4) Salt Lake City, (5) Atlanta.

Placing the parentheses at the right seems more natural and is probably less confusing to the pupil. Most persons find it easier to score responses in that position also but there are individual differences. Either method is preferable to having pupils underline the correct response.

2. It is generally better to have the alternate responses appear near the end of the item rather than at its beginning. Thus is it better to use the following form:

The commander of the United States land forces in the Mexican War was: (1) Lee, (2) Perry, (3) Mansfield, (4) Jackson, (5) Taylor

than to write it thus:

(1) Lee, (2) Perry, (3) Mansfield, (4) Jackson, (5) Taylor was the commander of the United States land forces in the Mexican War.

3. Be careful that the form of the question does not give a clue to the answer.

Poor: A pansy is a: (1) animal, (2) flower, (3) insect, (4) apple.

Better: A pansy is: (1) an animal, (2) a flower, (3) an insect, (4) an apple.

4. Avoid giving clues to the correct response by consistent use of a longer response as the correct one or by use of a shorter response for the correct one.

5. Use at least four responses, each having some plausibility. This suggestion presupposes that there are at least three plausible "foils." If this is not the case, there is no point in adding additional responses merely for the sake of uniformity. Unless one wishes to correct for chance there is no need of holding to a uniform number of responses. The danger in adhering to a fixed number is that there are likely to be a number of non-functioning responses which are used simply to secure the proper number.

6. Do not use the Multiple Choice technique if another form of test will serve your purpose better. Where there is only one correct response it is better to use the Recall type of item, particularly if the response is a single word or number. Where there are only two possible or plausible responses, the True-False or Yes-No items can be employed to better advantage.

7. Vary the order of the correct response so that it is not in the same position more than two or three times in succession. Unless one is on guard, it is easy to fall into the habit of placing the correct response in the same position in much too large a number of cases.

8. Do not use more than seven responses. To use a larger number complicates the reading problem unduly without contributing anything to the value of the test.

9. Remember that Multiple Choice questions can be used to test the pupils' judgment as well as their in-

formation but that it will accomplish this objective only in case provision is made in the questions. Consider the following item:

The main reason that the Mormons decided to settle where Salt Lake City now stands was: (1) There was an abundance of rainfall. (2) They wished to be undisturbed by other persons. (3) They did not like the mountains. (4) The country was easily reached by their friends.

For the pupil who has read about the Mormons but who has not been told why they chose this particular location, the above item may be a good exercise in inferential judgment. On the other hand, for a pupil who has read or who has been told of the reasons for settling in this location the selection of the correct response may be simply a matter of recall, of rote learning. It is true of all types of items, both of the objective and the essay types, that a memory question for one pupil may involve considerable reasoning for another.

The suggestions one might make for the Multiple Response items in which the pupil makes a plural choice are not essentially different from those given above.

An interesting variation in Multiple Choice and Multiple Response items is the practice of including a varying number of correct responses. For example, item number 1 might have three correct responses; item number 2, only one correct response; while number 3 might have two or four correct responses. Just what effect such procedure would have on pupil attitude and on chance and guessing factors is difficult to predict. Research in the usefulness of such tests have not been made to any extent and their possibilities need to be studied much further.⁸

⁸ Scheidemann, Norma V., Multiplying the Possibilities of the Multiple Choice Form of Objective Question, *Journal of Applied Psychology* 17:337-340, June 1933.

Recall Completion Tests

Recall Completion tests may be arranged in sentence form, ordinarily with a single blank in each sentence, or in paragraph form such as the following:

The colony was settled in 1607 in what is now the state of The leader of the settlers was Captain who was a leader. He was succeeded by Sir who made the colonists work like slaves.

Partly because of the difficulty of constructing good tests in paragraph form, the sentence type is more frequently used. As in the case of other tests, provision should be made for ease and rapidity in scoring. Since the blanks cannot always come at the end of the sentences, it will be found helpful to have pupils write their responses on designated blanks at the right of the page. These may be given the same numbers as those found in the blanks so as to guide the pupil.

Example: Zoology is a study of (1) life. (1)
 The tulip is a member of the
 (2) family. (2)

In constructing items of this type the following suggestions may prove helpful:

1. Avoid a large number of blanks in a single sentence.
2. Do not copy sentences directly from the text. To do so places a premium upon rote learning and photographic memory which is not the end sought in most instruction.
3. Avoid making blanks of a length to correspond with the length of the word. Make all blanks of uniform length. Avoid indicating the length of the word

by indicating a dot for each letter. This procedure is defensible in a test of intelligence where the emphasis is on verbal facility but not in a test of achievement.

4. Do not leave too many blanks in a single sentence. To leave several blanks confuses the pupil and makes scoring difficult.

5. If some other person is to do the scoring, see that the scoring key contains all of the possible correct responses. This is often difficult since it is the experience of most test makers that a surprising number of correct or near-correct responses occur to pupils which did not occur to the examiner.

6. Avoid supplying confusing articles or extraneous clues to the correct response. One of the sentences in the paragraph above reads: "The leader of the settlers was Captain who was a leader." The pupil who thinks of the word "able" for the second blank will be forced to discard it even though it be as suitable as any other word he can supply.

7. Be sure to leave enough space for the pupil to write so that his responses will not be illegible.

8. Avoid the possibility of one part of an exercise supplying or suggesting the responses required in other parts.

Matching Exercises

This type of test item is essentially a Multiple Choice test in which the same responses are employed for all items. The form indicated in the previous chapter (page 51) is satisfactory with slight modification for various types of items. The following suggestions should be borne in mind when constructing items of this sort.

1. Avoid having dissimilar items in any one series so that clues are given to the proper pairing. The following illustration from a teacher-made test will serve to make this warning clear.

- | | |
|-----------------------------|--|
| 1. Congo region | world's chief wool producer |
| 2. cog | great money crop of Egypt |
| 3. fluctuations in rainfall | hot rainy land |
| 4. tobacco | is practiced in delta of the Nile |
| 5. cotton | country with very regular coastline |
| 6. manioc | islands made up of coral and shell |
| 7. year-round cultivation | chief risk in crop raising in Argentina |
| 8. Africa | used for making bread in Amazon region |
| 9. Australia | railroads used on steep grades |
| 10. Bahama | grown in Cuba |
| 11. nitrate | is a source of wealth for Chile |

2. Do not include a very large number of items in either of the columns to be matched. The optimum number is probably ten or twelve, depending somewhat upon the maturity of the pupils and the subject matter involved. A larger number consumes much time and involves much reading and re-reading.

3. To decrease the chance factor, a larger number of items may be used in one series than is contained in the other. If the same number is used, the pupil who knows most of the answers can work out one or two by elimination.

4. Another method of decreasing chance is that of using the same response more than once. Pupils should be warned if this technique is used since they will otherwise feel that a response once used cannot be correct a second time. Directions should always be clear and complete.

5. If one list to be matched consists of names or words, arrange them in alphabetical order.

6. The use of matching exercises is generally fea-

sible only with relatively large units of material. Where only three to five items can be matched, it is probably better to use the Multiple Choice test which ordinarily takes less time to prepare and is likely to be less confusing to pupils.

A number of variations in constructing test items similar to those mentioned above have been reported. Particularly with the True-False tests there have been attempts at improvement. Several studies report experiments in which pupils were not only required to indicate whether statements were true or false but also to know enough about it to correct a false statement when they see one. This usually involves crossing out a word in each false item and substituting another word which makes the statement true.⁴ Andruss, for example, by unusual care in constructing his items, secured objective scoring and combined the advantages of the True-False, Multiple Choice, and Completion tests. Scoring was done by assigning one point for indicating that the item was false, one for crossing out the right word, and one for putting the correct word in its place.

Further Considerations

Which Type Shall the Teacher Use? The answer to this question depends upon a number of considerations. In the first place, the teacher must consider the purpose of testing. For example, does she wish to test the pupils' ability to recall or their ability to recognize? If the former, she will naturally choose the Recall Completion type; if the latter, she will choose one of the other types, since all of the other types discussed above involve recognition rather

⁴ Andruss, Harvey A., True and False Correction Test, *Balance Sheet* 16:61, October 1934.

than recall. Again, if the teacher's purpose is to diagnose pupils' individual difficulties so as to be able to apply remedial instruction, she will do well to avoid questions in which guessing or chance plays a large part. For example, if a pupil answers a True-False question correctly, the teacher cannot be certain whether the correct response was due to chance or to knowledge.

A second consideration is that of the type of question which can best be used with the type of material at hand. In some instances the material can best be covered by True-False or Yes-No questions; for other materials, the Multiple Response type is better, etc.

Still another consideration is that of the reliability of the various types of items. Information on this point is available from a number of studies by Charles, by De Graff and Ruch, and others, and may be briefly summarized as follows: For the same amount of working time the Recall Completion test is most reliable, except that well-constructed matching exercises of ten to twenty pairs probably rate a little higher. Next in order come the Single Choice seven-response or five-response types, while the True-False test has the lowest reliability.

The teacher will also, of course, be governed by her own preference for the various types.

Shall More Than One Type of Item Be Used in a Single Test? Statistical data with which to answer this question are lacking. However, it would seem that sufficient variety to relieve monotony would be desirable. If, on the other hand, one introduces too much variety, there is the danger of confusing the pupils and necessitating an undue consumption of time in the reading of directions. As pupils become more familiar with the various types of items, it may be desirable to increase the number of types in a given test. In general, it seems that for a 100-point test two

to five types of questions can be employed to advantage.

Should Pupils Be Instructed to Guess When in Doubt? Experiments by De Graff and Ruch and other investigators fail to throw definite light on this problem. It does appear, however, that the reliability of a test is increased if instructions against guessing are given. This appears to be especially true if the scores are corrected for chance.

Should Tests Be Corrected for Chance? The general formula employed where such corrections are made is the following: $\text{Score} = \text{Rights} - \frac{\text{Wrongs}}{N - 1}$, where N refers to the number of possible responses that the pupil can make. In the case of the True-False test or any two-response test, this formula becomes $\text{Score} = \text{Rights} - \text{Wrongs}$. In the case of two-response tests it is probably best to correct for chance since most of the studies indicate that scores on these tests are more valid and reliable than scores obtained on the same tests by counting the number of right responses. Ruch ⁵ suggests that the teacher who does not wish to go to the trouble of correcting for chance may make her tests as valid and reliable by making them ten or fifteen per cent longer than she had originally planned. There appears to be little value in correcting for chance scores on Multiple Choice items with four or more variants.

Formal and Informal Tests. The objective type questions discussed above are commonly used in standardized tests and also by teachers who construct their own objective examinations. The term "informal test" applies to both the objective and the traditional or essay test. Formal or standardized tests are usually more carefully constructed than the informal tests made by the teacher. They have

⁵ Ruch, G. M., *The Objective or New-Type Examination*, Scott, Foresman and Company, 1929, p. 356.

norms on the basis of which the teacher can evaluate the results; are generally less limited in scope; and usually have demonstrated reliability. More study has ordinarily been made also of the value of the individual items in a standardized test. Neither type should be used as the goal of instruction, since to be satisfied with results that compare favorably with the norms would be to admit that further improvement in the teaching process is either impossible or not to be desired.

While the author favors, in general, the new-type questions for informal examinations, it must be admitted that well-constructed essay examinations, scored in accordance with specific rules in mind, are often better than poorly-constructed objective tests. Either type must be constructed with care. The objective type requires, ordinarily, much more time for preparation, whereas the essay type usually requires much more scoring time.

May Objective Tests Be Presented Orally? It appears that the best method of administering objective tests is by having a copy of the test placed in the hands of each pupil. In this way the pupil may re-read and study certain items which on first presentation do not seem entirely clear. This procedure does involve some labor and even with the cheaper methods of duplication, the expense is an item to be considered in some school budgets. For these reasons there has been some experimentation with oral presentation of objective tests. While in general the practice of reading the tests has not seemed to lower either their validity or reliability, it will be noted that these experiments have been conducted almost exclusively with high school and college students. It is not known, therefore, just what the results would be if this mode of presentation were used in the lower grades. Probably shorter

items can be read without much harm. The writer has on several occasions used oral presentation of objective tests but feels that if for no other reason than the satisfaction gained by the pupil, it is better to place a copy of the test in his hands if possible.

PROBLEMS

1. Construct an essay test in some subject and write out rules for scoring which will make the test as objective as possible. If you can do so, administer the test, score the papers yourself but do not indicate your scores on the test papers. Have one or more other persons grade the papers and see how closely you agree.
2. Secure an informal True-False test or, if such a test is not available, a standard test containing True-False items. Note whether any of the suggestions made in this chapter have been violated. Are the violations justified? Why?
3. Construct an objective test covering Chapters I, II, and III of this text. Use at least three different types of items.
4. The following items have some faults. Indicate what is wrong and suggest means of improvement.

True—False

1. The Amazon River drains more land and carries more water to the sea than any other river.
2. Bananas are cut while they are still green.
3. We are not dependent upon the rest of the world for many commodities that we do not produce ourselves.
4. The defeat of Burgoyne which occurred in 1778 was considered the turning point of the American Revolution.
5. Truck farming is important in Florida and New York.
6. France hated England and wished to see her defeated.
7. A large sum of money was paid by the United States for Alaska.

Multiple Choice

1. Much wine is made in: (1) Bordeaux, (2) Lille, (3) Paris, (4) Lyon, (5) Calais.
2. The best method of purifying water is by: (1) the use

of chemicals, (2) letting it settle, (3) boiling, (4) freezing.

3. Many tulips and hyacinths are raised in: (1) Germany, (2) France, (3) Netherlands, (4) America.
4. The Frenchman who was of most help during the American Revolution was: (1) Kosciusko, (2) Lafayette, (3) von Moltke, (4) Pilsudski, (5) Pasteur.

BIBLIOGRAPHY

- Andruss, Harvey A., True and False Correction Test, *Balance Sheet* 16:61, October 1934.
- Briggs, T. H., and Armacost, G. H., Results of an Oral True-False Test, *Journal of Educational Research* 26:595-96, April 1933.
- Class, E. C., The Effect of the Kind of Test Announcement on Students' Preparation, *Journal of Educational Research* 28:358-61, January 1935.
- Frutchey, F. P., Measuring the Ability to Apply Chemical Principles, *Educational Research Bulletin* 12:255-60, December 1933.
- Greene, H. A., and Jorgensen, A. N., *The Use and Interpretation of Elementary School Tests*, Longmans, Green and Company, 1936.
- Hawkes, H. E., Lindquist, E. F., and Mann, C. R., *The Construction and Use of Achievement Examinations*, Houghton Mifflin Company, 1936.
- Hevner, Kate, A Method of Correcting for Guessing in True-False Tests and Empirical Evidence in Support of It, *Journal of Social Psychology* 3:359-62, August 1932.
- Horn, Helen R., The Variable Answer Test, *English Journal* (H.S. Ed.) 23:223-25, March 1934.
- McClusky, H. Y., The Negative Suggestion Effect of the False Statement in the True-False Test, *Journal of Experimental Education* 2:269-73, March 1934.
- Rinsland, H. D., *Constructing Tests and Grading in Elementary and High School Subjects*, Prentice-Hall, 1937.
- Ruch, G. M., *The Objective or New-Type Examination*, Scott, Foresman and Company, 1929.
- Scheidemann, Norma V., Multiplying the Possibilities of the Multiple Choice Form of Objective Question, *Journal of Applied Psychology* 17:337-40, June 1933.

Sims, V. M., and Knox, L. B., The Reliability and Validity of Multiple Response Tests When Presented Orally, *Journal of Educational Psychology* 23:656-62, December 1932.

Tiegs, E. W., *Tests and Measurements for Teachers*, Houghton Mifflin Company, 1931.

•

COMMON STATISTICAL TERMS AND PROCEDURES

•

Chapter IV

COMMON STATISTICAL TERMS AND PROCEDURES

MODERN EDUCATIONAL LITERATURE, dealing not only with testing but also with all sorts of experimental studies, is so replete with statistical terms and discussion of statistical devices, that the reader who does not have at least a passing acquaintance with statistical treatment of educational data is quite at sea. Furthermore, the user of tests who is to secure most help from these measuring instruments will need to manipulate the scores in order to make them most meaningful to herself, to her pupils, and to school authorities or patrons. It is the purpose of this chapter to present a few of the more common terms and techniques in order to help the teacher in her use of tests and so that later discussion of measurement may be more clearly understood.

Tabulating Test Scores and Other Data

When any considerable number of scores are gathered and are to be interpreted, it is customary to arrange them in a frequency table. This is not essential if the number of cases is small, but is usually helpful if one has a larger number of cases. A fifth grade class made the following scores on a forty-word spelling test: 29, 27, 25, 24, 28, 29,

40, 22, 32, 27, 33, 38, 20, 36, 31, 31, 25, 24, 21, 23, 29, 28, 31, 30, 39, 22, 29, 28, 33, 35, 34, 27, 36, 28, 27, 24, 32, 35, 37, 30, 31. Scores presented in this manner are difficult to interpret. By inspection we note that one pupil made a perfect score (40) while another made a score of 20, the lowest in this particular test. What of the others? Arrangement in a frequency table will enable us to see at a glance how the other scores are distributed.

Our first concern in devising a frequency table is to decide upon the size of the class interval. Since for ordinary work it is convenient to work with only ten or fifteen groups, we may decide in this case upon a two-point interval. If we accept what has been called the "mathematical" concept of number, that is, that 10 means from 9.5 to (but not including) 10.5, our table might appear as indicated in Table I.

TABLE I

Scores Made on a Forty-word Spelling Test

<i>Score Limits</i>	<i>Tallies</i>	<i>f</i>
39.5-41.5	/	1
37.5-39.5	//	2
35.5-37.5	///	3
33.5-35.5	////	4
31.5-33.5	///	3
29.5-31.5	/// /	6
27.5-29.5	/// ///	8
25.5-27.5	////	4
23.5-25.5	/// /	5
21.5-23.5	///	3
19.5-21.5	//	2
		<hr/> 41

In entering scores in a frequency table it will be convenient to tie the four previous tallies together by means of the fifth as indicated in the above table. Score limits

have been indicated in fractions in order to call attention to the usual method of considering a whole number. They are often indicated simply as 40-41, 38-39, etc.

The advantage of a frequency table consists not only in the fact that it enables one to see more clearly at a glance how the scores distribute themselves but also in the ease with which certain commonly used measures may be computed from it. This use will become more evident as the discussion proceeds.

Measures of Central Tendency

One of the first questions which is likely to be asked about a group of scores is: What is the average or central tendency? Three indices of central tendency are commonly employed.

The Mode. Probably the least satisfactory of these but the most easily found is the mode, which may be defined as the score appearing the largest number of times; that is, the most common score. In Table I we find that scores 28 and 29 occur most frequently; hence, we might say that the modal score is either 28 or 29. It will be noted that this measure of central tendency will be reasonably accurate only in case there is a tendency for the scores to center about the mid-point and usually only in case this distribution is fairly symmetrical. It does sometimes happen, however, that the mode rather than any other index of central tendency gives us the most accurate picture. If, for example, one were to ask, "What is the average income of the business men in your city?" the mode would probably give the most adequate picture of what the typical man earns. If there are some very large incomes, these would tend to raise the average and lead one to believe that the "typical" income is much higher than it really is. With

very large groups of data the mode is usually a rather reliable index because of the tendency for large groups of data to cluster about the mid-point.

The Median. A more commonly used measure of central tendency is the median which may be defined as the point in a distribution above which and below which lie an equal number of cases. Let us suppose that a class of seventeen pupils made scores on a 100-point test as follows: 88, 87, 85, 84, 82, 80, 80, 79, 78, 77, 76, 75, 73, 71, 68, 65, 61. The middle score is 78, since there are eight scores above it and eight below. In other words, 78 is the median, or, as it is sometimes called when obtained from ungrouped scores, the mid-measure. Had there been one more score, for example 60, the mid-point would have fallen between 77 and 78 and we would call the median 77.5. Where more cases are involved, it is customary to arrange the data in a frequency table and to compute the median by the process of interpolation, as indicated in Table II below.

TABLE II

<i>Score Limits</i>	<i>f</i>
80-84.99	2
75-79.99	5
70-74.99	5
65-69.99	9
60-64.99	6
55-59.99	4
50-54.99	4
45-49.99	2
40-44.99	1
	$N = 38$

Solution:

Divide the number of cases by two ($38 \div 2 = 19$). It is now evident that 19 cases are to be found in each half.

Adding up the *f* column we have 17 cases when the score 65 is reached. To find 19 cases we must take 2 out of the 9 cases in the next interval; that is, $2/9$ of the next class interval (which consists of 5 points) are needed. ($2/9$ of $5 = 10/9$ or 1.1.) This means that we need to go 1.1 points beyond 65, so we add 1.1 to 65 and find 66.1 as the median score.

As a check we may count down from the top of the *f* column. When we reach the interval 65–69.99 we have 12 cases; so 7 more are needed. ($7/9$ of $5 = 35/9 = 3.9$.) This means that we must go 3.9 points below 70 in order to find the median. ($70 - 3.9 = 66.1$, the median.)

The Mean. The central tendency which “the man on the street” would call the average is designated by statisticians as the arithmetic mean in order to distinguish it from other measures of central tendency. The procedure employed in finding it in the case of ungrouped data is sufficiently familiar to all. It consists in adding the raw scores and dividing by the number of scores. For example, the sum of 2, 3, 4, 5, and 16 is 30 and since there are 5 numbers the average or mean is 6. If one has scores arranged in a frequency table, the procedure is essentially the same except that one must multiply the mid-point of each interval by the number of cases at that interval. The procedure is indicated in Table III.

The scores in the above table represent the “rights” on an addition test. The total of the products column therefore represents the number of “rights” for all of the forty-three pupils. As will be seen in the table, this total is obtained by multiplying the mid-points of the various intervals by the frequencies of the intervals and adding these products. To compute the mean, we divide this total by the number of pupils which gives us $1392 \div 43 = 32.37$.

TABLE III

Finding the Mean from a Frequency Table (Long Method)

<i>Score Limits</i>	<i>Mid-Points</i>	<i>f</i>	<i>Products</i>
47-49	48	1	48
44-46	45	3	135
41-43	42	4	168
38-40	39	5	195
35-37	36	5	180
32-34	33	4	132
29-31	30	7	210
26-28	27	4	108
23-25	24	6	144
20-22	21	2	42
17-19	18	1	18
14-16	15	0	0
11-13	12	1	12
		$N = \overline{43}$	$\overline{1392}$ Total

TABLE IV

*Finding the Mean from a Frequency Table
(Short-Cut Method)*

<i>Score Limits</i>	<i>Mid-Points</i>	<i>f</i>	<i>d</i>	<i>f . d</i>
47-49	48	1	+6	+6
44-46	45	3	+5	+15
41-43	42	4	+4	+16
38-40	39	5	+3	+15
35-37	36	5	+2	+10
32-34	33	4	+1	+4
29-30	30	7	0	
26-28	27	4	-1	-4
23-25	24	6	-2	-12
20-22	21	2	-3	-6
17-19	18	1	-4	-4
14-16	15	0	-5	0
11-13	12	1	-6	-6
		$N = \overline{43}$		-32

The method presented above would prove rather cumbersome if we had several hundred cases instead of forty-three. A shorter method, sometimes called the short-cut method, has, therefore, been devised. The same frequency table is used as an illustration of this method (Table IV).

In the above illustration we have assumed the mean to be thirty. It is not necessary to assume it at this point but it seems to be in the vicinity of the mean and it is a bit more convenient to assume the mean at the approximate center of the distribution. Whether we do or do not the correction will be such that the result will be the same. In the deviation column (d) we note how far each class interval deviates from this assumed mean, those deviations above 30 being marked positive and those below marked negative. This is because those above tend to raise the mean above the point assumed while those below tend to lower the mean below the point assumed. Note that the figures in the d column represent deviations in terms of the number of class intervals, not in terms of actual units. Adding the positive deviations multiplied by their frequencies we have 66. Adding the negative deviations multiplied by their frequencies we have 32. To find the correction on the assumed mean, we subtract the total negative deviations from the total positive deviations, multiply by the size of the class interval, and divide by the number of cases. Thus we have $\frac{66 - 32}{43} \times 3 = + 2.37$. This correction added to the assumed mean gives us $30 + 2.37$ or 32.37, which is the same result as that obtained by the longer method. The distinct advantage of this method is that we work with smaller numbers.

Measures of Variability

Although a measure of central tendency gives us important information concerning a series of data, there are other important things to know about the distribution. We are usually interested in the spread or scatter of the scores as well as their average. One indication of the spread is the *range*, or the distance from the lowest to the highest score. The range is, however, only a very rough measure of dispersion. If, for example, we are told that the range of a series of scores is from 12 to 93, we might infer that the scores are evenly or symmetrically distributed between these two points. While this may be true it is also possible that there is only one score below 50 and that this score is 12. In order to give a more adequate indication of the spread of scores, various measures of deviation are used, two of the simpler and more common of which will be taken up briefly.

The Semi-Inter-Quartile Range. This index of deviation is not so commonly employed, partly because it does not supply us with as much information as do some of the other measures of variability. It merits our attention, nevertheless, since the upper and lower quartiles on which it is based are very commonly indicated, particularly in reporting test norms.

Consider the following table of scores (Table V) on a test in American government.

To find the semi-inter-quartile range we must first find the lower quartile (Q_1) and the upper quartile (Q_3). To find Q_1 , proceed as follows: Divide N by 4 in order to find one-fourth of the cases. In this problem it is 15. Count up the f column until you come as near to 15 as possible without going beyond. This brings you to the interval labeled 55-59, the lower limit of which is really 54.5 Up to that point there are 11 cases, so you need 4 cases of the 5 in the

TABLE V
Scores on a Test in American Government

Scores	f
90-94	1
85-89	3
80-84	5
75-79	7
70-74	9
65-69	11
60-64	8
55-59	5
50-54	4
45-49	3
40-44	1
35-39	2
30-34	1
$N = 60$	

next interval. Take $\frac{4}{5}$ of 5, since 5 represents the size of the class interval. This gives you 4 which is to be added to the lower limit of the class interval; namely, 54.5. Hence the lower quartile point is 58.5.

To find Q_3 , count up the f column to get $\frac{3}{4}$ or 45 of the cases. This brings you to the interval of which the lower limit is 74.5. Since there are 44 cases up to this point, we see that we must have 1 of the 7 scores at that interval. Proceeding as before, we have $\frac{1}{7}$ of 5 or .7 as the correction to be added to 74.5. The upper quartile point is thus 75.2. The semi-inter-quartile range (Q) is obtained by subtracting Q_1 from Q_3 and dividing by 2, so in this case we have $\frac{75.2 - 58.5}{2}$ or 8.35. This means that if we go below the median 8.35 points and above it by the same amount we would expect to include about 50 per cent of the cases. The median for this table is found to be 67.2. Adding to this 8.35 we have 75.55; subtracting we have 58.85. Between 58.85 and 75.55 then, we expect to

find about 50 per cent of the cases. In actual practice we find some slight variation from 50 per cent unless the distribution is quite symmetrical.

The Standard Deviation

The most common measure of variability used in educational data is the standard deviation, the symbol of which is the Greek letter sigma (σ). One method of computation is indicated in Table VI.

TABLE VI

A Method of Finding the Standard Deviation

<i>Scores</i>	<i>f</i>	<i>d</i>	<i>d</i> ²	<i>fd</i> ²
24	1	+5	25	25
23	0	+4	16	0
22	2	+3	9	18
21	5	+2	4	20
20	6	+1	1	6
19	9	0	0	0
18	6	-1	1	6
17	4	-2	4	16
16	1	-3	9	9
15	2	-4	16	32
	<u>36</u>			<u>132</u>

Since the mean of these scores is 19, the number 24 deviates from it by +5, 23 by +4, and so on. Square these deviations as shown in the d^2 column; then multiply by the frequencies to secure the fd^2 column. The standard deviation is now found by dividing the sum of the fd^2 column by the number of cases and extracting the square root. In our illustration we thus have $\sqrt{132 \div 36}$, or 1.9. Note that this measure is computed from the mean as the central tendency. When it is laid off on each side of the mean, it includes about two-thirds of the cases. Thus in

the above illustration we expect to find about $2/3$ of the cases between 20.9 and 17.1.

A shorter method of finding the S.D. for grouped data is indicated in Table VII.

TABLE VII
Finding the S.D. by the Short-Cut Method

Scores	f	d	fd	fd ²
90-94	2	+3	+6	18
85-89	5	+2	+10	20
80-84	8	+1	+8	8
			+24	
75-79	9	0	0	0
70-74	6	-1	-6	6
65-69	3	-2	-6	12
60-64	3	-3	-9	27
55-59	3	-4	-12	48
50-54	1	-5	-5	25
			-38	
	<u>40</u>			<u>164</u>

As we did when we found the mean by the short-cut method, we begin by assuming a mean. In this case we have chosen 77, the mid-point of the interval 75-79. In the d column we indicate the deviations of each interval from the interval containing the assumed mean. Multiplying the f's times the d's we have the fd column, from which the correction on the assumed mean may be computed as well as the correction of the S.D. The formula for the standard deviation is: $\sigma = s \sqrt{\frac{\sum fd^2}{N} - c^2}$ in which $\sum fd^2$ is the sum of the fd² column, c the size of the correction, and s the size of the interval. To find c we take the algebraic sum of the fd column and divide by the number of cases. Thus $c = \frac{+24-38}{40} = \frac{-14}{40}$ or $-.35$ $c^2 = +.1225$.

Substituting in the formula we have $\sigma = s \sqrt{\frac{164}{40} - .1225}$
 $= s \sqrt{3.9775} = s \times 1.99$. The size of the class interval being 5, we multiply by this figure and obtain 9.95.

Many students find difficulty in comprehending the meaning of the S.D. Remember that it is a measure of the spread of scores and may be considered much as a yardstick except that it varies in size with each new set of data.

Measures of Relationship

In the interpretation of test results it is very often necessary to know not only about the central tendencies of scores and their variability, but also to understand the relationship of two series of scores or two sets of data. The most common index of the relationship between two sets of data is the coefficient of correlation. The coefficient of correlation has been defined as an index of the degree of "going-togetherness" of two sets of scores. Let us suppose, by way of illustration, that seven pupils made scores as follows on forms A and B of a given test.

<i>Pupil No.</i>	<i>Form A</i>	<i>Form B</i>
1	92	83
2	85	87
3	62	71
4	87	62
5	60	91
6	49	85
7	80	82

Since the two forms of the test are supposed to be of equal difficulty, one would normally expect that a pupil who scored high on one form would also do well on the second. In the above case, however, one could not make any accurate prediction of score on form A if one knew the

score on form B. It is not to be understood that the above represents a typical case since ordinarily the scores on two forms of a test do correspond much more closely than is there indicated. It is the function of the coefficient of correlation to show how close the relationship is. If, for example, the pupil who scores highest on one test also scores highest in the other and the relative rankings of all the other pupils are about the same on both tests, we should find that the coefficient of correlation would be high, that is, approaching $+1.00$. If, on the other hand, the rankings on one test were to be almost reversed so that the pupil rating highest on one rated lowest on the other, and so on throughout, the coefficient would again be high but negative; that is, approaching -1.00 . Again, if there is almost no relationship between scores on two tests, the coefficient would be low; that is, approaching zero.

One of the more common methods of finding the Pearson Product Moment coefficient of correlation is illustrated in a very short problem in Table VIII.

TABLE VIII

<i>Pupil</i>	<i>English Grade</i>	<i>History Grade</i>	<i>x</i>	<i>y</i>	<i>xy</i>	<i>x²</i>	<i>y²</i>
A	80	75	-4	-8	32	16	64
B	70	79	-14	-4	56	196	16
C	82	77	-2	-6	12	4	36
D	83	85	-1	2	-2	1	4
E	86	82	2	-1	-2	4	1
F	87	92	3	9	27	9	81
G	85	83	1	0	0	1	0
H	84	86	0	3	0	0	9
I	93	90	9	7	63	81	49
J	90	81	6	-2	-12	36	4
Total	840	830	0	0	174	348	264
Mean	84	83				34.8	26.4

$$r = \frac{\text{Sum } xy}{n \cdot \text{sigma } x \cdot \text{sigma } y} = \frac{174}{10 \times \sqrt{34.8} \times \sqrt{26.4}} = \frac{174}{300.9} = .578$$

The steps are as follows:

1. Find the mean of the English grades (84) and also the mean of the history grades (89).
2. Record in the x column the individual deviations from the mean of the English grades, being careful to observe the signs. For example, the score 80 deviates -4 from 84 and 70 deviates -14 . In column y , record deviations from the mean in history. The accuracy of the work up to this point may be checked by taking the algebraic sum of the x and y columns. This should always be equal to 0.
3. Multiply the x column by the y column, again noting signs. Take the algebraic sum of this xy column. This is the numerator in the formula.
4. Square the x column and record in a separate column. Do the same for the y column. Take the sum of these columns.
5. Sigma x is now found by extracting the square root of the mean of the x^2 column and sigma y is likewise found by taking the square root of the mean of the y^2 column.
6. Substitute in the formula and solve.

The Pearson Coefficient Computed from a Scattergram

It sometimes happens that one is interested in knowing whether any relationship exists between two sets of data, but does not care to determine the coefficient of correlation. Sometimes a mere inspection of the scores will reveal a correlational tendency but if there are a considerable number of exceptions in a large number of cases, one can easily be mistaken in the existence or non-existence of such tendencies. For this reason various graphic devices are sometimes used. For example, an instructor in Introduction to Education wondered whether or not there was any tendency for the students who made high scores on the freshman intelligence test to earn better grades than

did those students who made low intelligence test scores. To enable him to answer this question the grades assigned were divided into three groups; namely, those below "C," the "C" grades, and those above "C." Similarly, the freshman test record was examined to see which students had fallen in the lower one-third, middle one-third, and upper one-third in intelligence test scores. The position of each of the eighty-one students in both intelligence and grades was then indicated in a table like the one below.

	Below C	C	Above C
Upper $1/3$	1	9	13
Middle $1/3$	5	24	6
Lower $1/3$	15	7	1

It can now be seen that 15 students who were in the lowest one-third in their intelligence test scores received grades below "C"; whereas, only 1 student who was in the lowest one-third in intelligence earned a grade above "C." An examination of this table thus reveals a marked tendency for students to earn grades that are comparable to the test scores made. However, it will be noted that there are some exceptions.

It is also possible to compute the coefficient of correlation from the data as arranged in the above table. However, since the grouping is so crude, the results would probably not be very accurate. With a somewhat finer grouping, however, this method of determining correlation is convenient, particularly where the number of cases is large. The method is indicated in the accompanying correlation chart (page 98) prepared to show the rela-

Name of Test on Y-axis: A 100-point American History Test

Name of Test on X-axis: A 50-point American History Test																		
	15	18	21	24	27	30	33	36	39	42	45	48	f	d	f.d	f.d ²	$\sum x \cdot y$	$\sum x^2$
95	-30	-25	-20	-15	-10	-5		+5	+10	+15	+20	+25	4	+5	20	100	70	
90	-24	-30	-16	-12	-8	-4		+4	+8	+12	+16	+20	7	+4	28	112	76	4
85	-18	-15	-12	-9	-6	-3		+3	+6	+9	+12	+15	15	+3	39	117	75	12
80	-12	-10	-8	-6	-4	-2		+2	+4	+6	+8	+10	20	+2	40	80	56	14
75	-6	-5	-4	-3	-2	-1		+1	+2	+3	+4	+5	22	+1	22	22	25	7
70													32		+140			
65	+6	+5	+4	+3	+2	+1		-1	-2	-3	-4	-5	22	-1	-22	22	29	7
60	+12	+10	+8	+6	+4	+2		-2	-4	-6	-8	-10	15	-2	-26	52	56	8
55	+18	+15	+12	+9	+6	+3		-3	-6	-9	-12	-15	8	-3	-24	72	66	
50	+24	+20	+16	+12	+8	+4		-4	-8	-12	-16	-20	5	-4	-12	48	52	
f	2	5	7	11	14	17	27	20	18	12	8	5	144		-84	685	505	58
d	-6	-5	-4	-3	-2	-1		+1	+2	+3	+4	+5					451	
f.d	-12	-25	-28	-33	-28	-17	-145	20	36	36	32	15	+139					
f.d ²	72	125	112	99	56	17		20	72	108	128	75	884					

tionship between scores on a history test and scores on a test in geography.

Study the correlation chart to see how it is ruled. Notice that the scattergram part is laid off with respect to the X-axis (horizontal) and the Y-axis (vertical), which cross each other near the center. Notice that most cells contain a figure with a plus or minus sign. The figure indicates the size of the "product-moment." This is somewhat like the leverage illustrated by a teeter; one's effect in pulling the teeter downward depends on two things, one's weight and one's distance from the fulcrum. In the scattergram the fulcrum is the cell at which the two axes cross. The column headed "d" and the row headed "d" indicate the distances from the fulcrum; their product gives the effect which each cell is capable of producing.

The scores have been tallied into the table. Thus a pupil making a score of 95 (from 92.5 to 97.5) in history and a score of 48 (from 46.5 to 49.5) in geography is entered in the upper right-hand cell. After all the scores have been tallied, the rows are added to fill the f column; the columns are added to fill the f row. The remaining part of the work, except the two xy columns, is the process of finding the standard deviation, which you have had previously. Let us compute the upper row for the xy value. It is arrived at thus: $1 \times 25 + 2 \times 20 + 1 \times 5$ gives us 70, which we enter in the $+x \cdot y$ column, since all the signs of the cells were plus. To calculate the second row: $2 \times 20 + 1 \times 16 + 1 \times 12 + 1 \times 8$ gives us 76, which we enter in the $+x \cdot y$ column. The case 1×-4 gives us a -4, which we enter in the $-x \cdot y$ column. The case which appears on the Y-axis we take no note of, its cell value being 0. It is like a person on the teeter exactly over the fulcrum; his weight contributes nothing to either end. We do all the rows in like fashion. Then we find the algebraic

total of the $x \cdot y$ columns. We are now ready to apply the formula for finding r :

$$r = \frac{\frac{\sum x \cdot y}{N} - c_x \cdot c_y}{\sigma_x \cdot \sigma_y}$$

Notice that this formula involves the standard deviation of each set of scores and the amount of correction for the assumed mean for each set of scores. These corrections are designated by c_x and c_y , indicating the correction obtained from the lower row f.d for the X-axis (Geography) and the column f.d for the Y-axis (History). Likewise σ_x (sigma sub x) represents the standard deviation for the geography scores and σ_y represents the standard deviation for the history scores. (All in terms of class intervals.) Computations:

$$c_x = \frac{-143 + 139}{144} = -.03 \quad \sigma_x = \sqrt{\frac{884}{144} - .0009} = 2.47$$

$$c_y = \frac{-84 + 149}{144} = +.45 \quad \sigma_y = \sqrt{\frac{625}{144} - .2025} = 2.03$$

$$c_x^2 = .0009$$

$$\sigma_x \cdot \sigma_y = 5.0141$$

$$c_y^2 = .2025$$

$$\frac{451}{144} - (-.0135)$$

$$c_x \cdot c_y = -.0135$$

$$r = \frac{5.0141}{5.0141} = +.63$$

If one wishes to find the correct standard deviation, one must multiply by the size of the class interval so that S.D._x becomes 3×2.47 , or 7.41, and S.D._y becomes 5×2.03 , or 10.15. In determining the mean, combine the correction (multiplied by the size of the class interval) with the assumed mean. Thus:

$$M_x = 33 - (3 \times .03) = 32.91,$$

and

$$M_y = 70 + (5 \times .45) = 72.25$$

Some of the uses of the coefficient of correlation will be discussed in connection with the reliability of tests, in Chapter XII.

PROBLEMS

- The following scores were made on a one-hundred-point civics test: 73, 61, 85, 70, 79, 58, 72, 64, 67, 81, 59, 48, 66, 92, 74, 71, 64, 82, 76, 55, 62, 39, 54, 70, 71, 62, 56, 91, 67, 84, 78, 51, 68, 83, 74, 46, 72, 58, 79, 75.
 (A) Tabulate these scores into a frequency table, using a five-point class interval, the lowest interval being 35-39.
 (B) Find the mean of the above scores by the short-cut method. (Remember that a score of 35 means from 34.5 to 35.5, etc.)
 (C) Find the mean by adding all the scores above and dividing by N. By how much does this mean differ from the mean as found by the short-cut method? How do you account for this difference?
- Find the median for the following table. Then condense the table into one in which the size of the class interval is 20; that is, where the lowest interval is 0-19.5, etc. Compute the median with this arrangement, noting whether or not the result differs from that obtained in the five-point table.

<i>Scores</i>	<i>f</i>
94.5-99.5	2
89.5-94.5	3
84.5-89.5	2
79.5-84.5	4
74.5-79.5	6
69.5-74.5	5
64.5-69.5	6
59.5-64.5	9
54.5-59.5	10
49.5-54.5	14
44.5-49.5	10
39.5-44.5	7
34.5-39.5	5
29.5-34.5	7
24.5-29.5	5

<i>Scores</i>	<i>f</i>
19.5-24.5	7
14.5-19.5	4
9.5-14.5	4
4.5- 9.5	6
0 - 4.5	2

3. Compute the median and the value of Q in the following table.

<i>Scores</i>	<i>f</i>
95-99	2
90-94	1
85-89	5
80-84	8
75-79	7
70-74	7
65-69	5
60-64	3
55-59	1
50-55	1

4. Compute the mean and the S.D. of the scores below, using the short-cut method.

<i>Scores</i>	<i>f</i>
140-149	2
130-139	5
120-129	10
110-119	19
100-109	26
90- 99	15
80- 89	12
70- 79	6
60- 69	4
50- 59	1

5. Compute the coefficient of correlation for the following sets of scores by the method indicated in Table VIII.

<i>A</i>	<i>B</i>
3	25
85	50
55	37
39	18

<i>A</i>	<i>B</i>
95	95
89	80
80	83
53	25
92	96
96	97
52	80
30	12
43	43
42	40
66	33

6. The following are scores made on an English test and on the Henmon-Nelson Tests of Mental Ability. Using the scattergram method, compute the coefficient of correlation.

<i>English</i>	<i>H-N</i>	<i>English</i>	<i>H-N</i>
170	29	181	33
173	34	193	50
119	22	146	32
189	49	186	43
124	21	192	42
194	45	167	39
202	48	179	36
117	33	180	32
201	42	191	42
174	44	161	22
180	20	145	46
183	39	172	28
101	20	132	29
171	41	196	37
135	37	174	33
147	23	151	45
167	46	115	19
199	43	197	42
96	36	169	26
175	42	151	31
195	52	186	43
195	40	208	41
183	21	193	39
194	54	139	31
156	36	131	15
155	36	204	44

<i>English</i>	<i>H-N</i>	<i>English</i>	<i>H-N</i>
147	23	189	47
171	36	187	39
165	27	165	16
192	39	191	29
134	24	168	39
137	37	152	45
161	31	192	39
193	47	185	37
190	36	165	36
155	26	195	34
163	41	193	32
166	34	181	34
170	29	153	32
186	39	180	48
215	50	128	40
200	45	158	31
148	30	210	55
176	30	187	50
176	28	141	32
176	53	97	13
182	55	184	44
100	26	196	41
176	46	176	34
161	31	107	22

BIBLIOGRAPHY

- Enlow, E. R., *Statistics in Education and Psychology*, Prentice-Hall, 1937.
- Garrett, H. E., *Statistics in Psychology and Education*, Longmans, Green and Company, 1926.
- Good, Warren R., *The Elements of Statistics*, The Ann Arbor Press, 1933.
- Holzinger, Karl J., *Statistical Methods for Students in Education*, Ginn and Company, 1928.
- Kramer, Edna E., *Educational Statistics*, John Wiley and Sons, Inc., 1935.
- Lindquist, E. F., *A First Course in Statistics*, Houghton Mifflin Company, 1938.
- Nelson, M. J., and Denny, E. C., *Workbook in Statistics for Teachers*, Iowa State Teachers College, Cedar Falls, Iowa, 1938.
- Sorenson, Herbert, *Statistics for Students of Psychology and Education*, McGraw-Hill Book Company, 1936.
- Wert, James E., *Educational Statistics*, McGraw-Hill Book Company, 1938.

•

TESTS IN READING AND ARITHMETIC

•



Chapter V

TESTS IN READING AND ARITHMETIC

Reading

IT WOULD BE wholly impossible to over-estimate the importance of reading as a school subject or as an activity outside of the schoolroom. What the child learns in history, geography, arithmetic, or other school subjects is dependent in no small measure upon the fluency with which he reads and upon his comprehension of the printed page. Outside of school his reading ability is utilized many times daily in interpreting traffic notices, reading advertisements, informing himself concerning current happenings, locating business establishments, selecting food. One has but to list the activities for a brief time to become conscious of the demands which are made upon one's ability to interpret printed or written matter. Printed materials may, moreover, become a tremendous force in molding the thinking of a nation. No doubt the press exerts a very considerable influence on political and economic thought and on every phase of modern civilization. Yet, this fundamental tool-subject of reading, because of its great complexity, offers many and serious obstacles to the expert in the field of educational measurement.

The earlier tests in this field concerned themselves with an attempt to determine the rate of oral or silent reading

or the amount of comprehension in one of these divisions of reading, or both the rate and the comprehension. It must be noted that no direct measure of a pupil's comprehension can be obtained. The best we can do is to ask the pupil to perform some supplementary task such as answering questions concerning the material which he has read. This procedure involves making the test situation somewhat artificial but perhaps it appears less artificial to the child than one might imagine since the present-day child is required to respond in much the same manner in numerous tests in the various school subjects, as a part of his everyday work.

Tests of Silent Reading. Some of the earlier tests in reading, such as the Monroe Silent Reading Test,¹ secured in a single test and in a very brief time a measure of both speed and comprehension. Such tests were not concerned with different types of reading, or at least did not make provision for indicating in detail a knowledge of the vocabulary used, ability to follow directions, or a complete understanding of context. Recent researches have revealed that there are a great many types of reading ability, and that ability to do one type of reading well does not always indicate ability to perform well in all types of reading. It is interesting to note how tests have been constructed to include more of the different types of materials and how provision has been made for securing separate scores for the different parts. Thus the Stanford Achievement Test in Reading² provides for the reading of paragraphs and for the understanding of vocabulary. The Gates Primary Reading Tests³ measure three types of reading; namely,

¹ Published by Public School Publishing Company.

² Published by World Book Company.

³ Published by the Bureau of Publications, Teachers College, Columbia University, New York.

word recognition; word, sentence, and phrase reading; and the reading of directions. The Gates Silent Reading Tests,⁴ Nelson's Silent Reading Test,⁵ and the Instructional Reading Tests for Intermediate Grades by M. J. Nelson,⁶ are so constructed that four separate measures are secured. More ambitious attempts have been made in the Sangren-Woody Reading Test ⁶ which has seven divisions and the Iowa Reading Test ⁶ which has six.

Some Fundamental Questions Concerning the Testing of Reading. One who examines tests in silent reading is confronted with a number of questions which have a direct bearing upon his choice of a test to use. One of these questions relates to the amount of time which is required for a fairly adequate measure of reading ability. While the Monroe Silent Reading Test requires but four minutes, there now seems to be rather general agreement, if one may judge by current practice, that so short a time is quite inadequate. Most of the tests appearing in recent years have devoted at least fifteen minutes to the measurement of reading and some of them much more. Because of the lack of reliability of shorter tests, the classroom teacher would probably be unwise in selecting an extremely short test. On the other hand an administrator wishing to make only group comparisons might find it desirable to use a very rapid survey instrument if the time for measurement is extremely limited.

How many different reading abilities may be measured in a single test of fifteen to thirty minutes? No definite answer can be given to this question but it appears that, if many subdivisions of a test are made, one will secure

⁴ Published by the Bureau of Publications, Teachers College, Columbia University, New York.

⁵ Published by Houghton Mifflin Company.

⁶ Published by World Book Company.

an inadequate measure of each. Furthermore, for the most part, there is no need of subdividing measurement in the field very minutely. While it might on the surface seem desirable to measure as many types of reading ability as possible, this is not necessarily the case. In fact, it may be much more desirable to secure a fairly adequate measure of certain "key" types of reading ability. In discussing this question, Gates⁷ says: "Although there are many types of reading techniques, it does not follow that it is necessary to measure every one for purposes of diagnosis. It is probable that a certain small number of tests may be found which indicate the significant strengths and weaknesses sufficiently well for practical purposes."

Two other questions of considerable significance relate to the method of arranging reading materials in tests. Shall the items be of about equal difficulty so that the pupil's score is dependent upon the number of items he can cover, as well as his comprehension of material of a given level? Or shall the items be of increasing difficulty so that the pupil's score becomes not only an index of the amount of material read, but also an indication of his ability to do more and more difficult reading tasks? The first arrangement is illustrated by the Gates Silent Reading Tests and the second by the Nelson Silent Reading Test. No dogmatic answer can be given to this question but the answer of the individual teacher to this question will naturally determine in part her selection of the test she will use. The second of these questions may be stated as follows: Is it better to arrange reading test items in such a way that all like items are together and the pupil becomes aware of the kind of question that will be asked in each case, or is it better for the test items to be mis-

⁷ Gates, A. I., *Improvement of Reading*, The Macmillan Company.

cellaneously arranged so that the pupil is not aware of the exact nature of the questions to be asked? The Gates and Nelson tests again may be used to typify the difference, since in the Gates tests one entire section is devoted to reading for the general significance of the paragraph, another section is devoted to the prediction of probable outcomes, and so on. In the Nelson test, on the other hand, the questions on general significance, details, and probable outcomes are miscellaneously arranged so that the pupil does not read for a single purpose. The question really resolves itself into this: Which is the more natural reading situation? Again no dogmatic answer can be given, for some will argue that the pupil ordinarily reads for a specific purpose, while others contend that the purpose is rarely so specific as that of noting only details, or only probable outcomes. Here again is a question which the individual teacher will need to answer to her own satisfaction.

Tests of Oral Reading. Measurement of skill in oral reading has been given much less attention than measurement in silent reading. This is the result, in part, of the greater attention which silent reading has received because of the discovery that more silent reading is practiced by the average person than oral reading, and that the measurement of skill in oral reading requires individual testing which is always more time-consuming. It must be admitted, however, that oral reading is of some importance and that considerable attention must be devoted to this type of reading, especially in the earlier grades. It is moreover true that on those occasions when one is required to do oral reading, it is frequently a source of keen embarrassment if one is not able to read well.

Gray's Oral Reading Check Tests^{*} are designed pri-

^{*} Published by Public School Publishing Company.

marily for diagnosis of difficulties but a composite score may be obtained and norms to which the scores may be referred are available for each grade. These norms are expressed in terms of rate of reading and in terms of errors made. The Gates Graded Word Pronunciation Test⁹ does not seek to determine fluency in running passages, but only ability to pronounce isolated words. Gates has established age and grade norms for this test and has listed the corresponding scores on Gray's Oral Passages, so that direct comparisons can be made of the pupil's performance on the two tests. As a means of diagnosis he suggests that both of these be used. "If the pupil does appreciably better on the Gray's Passages than on the Gates' Words, it is highly probable that the context has been fruitfully utilized. If the age or grade scores are approximately the same, it is probable that the context is utilized to an average degree. If the score is appreciably better on the Word Pronunciation Test, it is probable that the pupil makes relatively little use of the context—a deficiency that certainly should be remedied."¹⁰

Improvement of Reading. Much has been done in recent years to improve the reading ability of school children. Perhaps the most important change has been in the amount of reading material provided. Where formerly each pupil was given a single "reader" for each grade, the modern school provides a dozen or even more. It is thus not necessary for the pupil to read and re-read the same material time and again, which is always boring to many children. The wealth of material now available is not only conducive to better reading habits but is also much more instructive in other ways. While considerable

⁹ Published by the Bureau of Publications, Teachers College, Columbia University, New York.

¹⁰ Gates, A. I., *Improvement of Reading*, p. 136, The Macmillan Company.

improvement has thus been made, much can also be accomplished by careful attention to the special weaknesses of pupils as revealed by diagnostic tests. The tests described above will be of material assistance if carefully used.

It might appear on the surface that poor reading habits, which have been established in the lower grades, can be overcome only with extreme difficulty. That this is not the case, however, is evident from a considerable number of studies which indicate that even in the upper grades much can be accomplished by attention to the improvement of reading. Jacobson and Van Dusen¹¹ found, for example, that the poor readers in the ninth grade made large gains when they were: (1) encouraged to read much easy material; (2) given a considerable amount of work-type reading such as is found in the *Learn to Study Readers* by Horn and others; (3) assigned materials which developed paragraph comprehension by asking questions concerning the paragraph read. As might be expected, the pupils with higher intelligence, who were poor readers, were found to make the largest gains, but even the pupils of lower intelligence showed considerable improvement. As additional evidence that improvement is possible, Pressey and Pressey¹² found that those poor readers among college freshmen who had special training not only improved in reading, but also did better work in their academic subjects than did similar students without the training. As an indication of some of the remedial work which may prove to be effective in some instances, Stone¹³

¹¹ Jacobson and Van Dusen, Remedial Instruction in Reading in the Ninth Grade, *School Review* 38:142-146, February 1930.

¹² Pressey and Pressey, Training College Freshmen to Read, *Journal of Educational Research* XXI: 203-211, March 1930.

¹³ Stone, Clarence R., A Non-Reader Learns to Read, *Elementary School Journal* 30:142-146, October 1929.

lists the following remedies which were used to advantage with a boy of seven years and nine months (I.Q. = 87) who had attended the first and second grades for two years and was unable to read: (1) Simple materials which he would desire to read were placed before him. The Child's-Story Primer proved to be a wise choice. (2) Word recognition was developed by many typewritten exercises in large print. (3) He was taught to write from memory the words used in reading. (4) Special attention was given to the distinction between "d" and "b" and other letters which are similar and prove to be confusing to the child. (5) Daily drills in both meaningful content and lists of words were given. (6) Various types of matching exercises were used to develop ability to see likenesses and differences. (7) Care was taken to give some joyful experience in reading each day. (8) Some oral reading was provided each day, but the emphasis was mainly on the silent reading techniques.

A wealth of easy material is recommended, especially for duller pupils, if appropriate reading habits are to be established. It must be remembered that lack of interest in reading is perfectly natural among poor readers, for all of us are interested in reading only when we derive meaning from what we read.

Among the more persistent difficulties is the inability to comprehend the character of the word forms. Mirror readers show reversals in reading, such as reading "saw" for "was" or "on" for "no," while other children are confused by the similarity of certain letters, such as "m" and "n" or "d" and "b." Though not recommended for normal readers, pupils with these difficulties can be helped by calling their attention to separate letters and by having them pronounce words in syllables. Another method

sometimes used, but which should be used only with cases that fail to respond to other types of treatment, may be called the kinesthetic method.¹⁴ When the child fails to recognize a word, the word is written on a piece of paper and the child is directed to trace it with his finger, repeating the word slowly as he traces. When he thinks he has traced the word often enough, the copy is removed and the child prints or writes the word, pronouncing it again as he writes. This method is probably most effective when the child has learned print-like writing, since he is then not confused by the difference between script and printed material.

Teacher-Made Tests in Reading. It is probably true that more of the testing can be done to advantage by means of standard tests in the field of reading than is the case in many fields, but from time to time the teacher may wish to make short reading tests of her own. However, care must be taken to make the testing situations sufficiently intriguing so as not to destroy that most valuable of assets, the love of reading. Lest the pleasure in reading be diminished, it is probably unwise to test materials designed for pure enjoyment. There is grave danger, for example, in detailed testing with such stories as *Cinderella*, *Robinson Crusoe* and a host of others. In work-type reading, on the other hand, testing is often valuable. For this purpose the teacher may select paragraphs from work-type reading as it appears in a text and arrange suitable questions about them but often she will find it to her advantage to rewrite such materials in order that they may be better suited to her purposes. One of the pitfalls she should avoid is that of giving the answer to her questions in the very opening words of the paragraph.

¹⁴ For a full description of this method see Gates' *Improvement of Reading*, pp. 306 ff.

Consider, for example, the following selections from the Monroe Silent Reading Test.

3. Lincoln loved birds and animals. It hurt him to have any of them suffer. Even when he was very busy, he would stop to help an animal in distress.

Draw a line under the word which best describes Lincoln.

tall great angry wicked kind

4. Robins in the tree-top,
 Blossoms in the grass,
 Green things a-growing,
 Everywhere you pass;
 Sudden little breezes,
 Showers of silver dew,
 Black bough and bent twig
 Budding out anew.

Draw a line under the month which is described in this stanza.

April January August October December

It is evident that the pupil (particularly a bright child) does not need to read the entire paragraph in order to answer the questions. It may be argued that if the pupil is bright enough to respond in this way, he will also be a good reader, since there is a fairly close relationship between reading ability and intelligence. Such argument is fallacious, however, for in spite of this close relationship in general, there are many exceptions. Furthermore, in order for a test to be valid it must measure what it purports to measure, not something related to it, no matter how closely related.

Reading Readiness Tests. A comparatively new development in the field of testing is the publication of tests to determine whether pupils are ready to start their ex-

periences in reading. Capacity to learn to read has for some time been indicated by scores on intelligence tests and these tests of reading readiness have much in common with primary tests of intelligence. In addition to items typical of the usual intelligence test, however, one finds motor tests of various sorts, auditory tests, pronunciation tests, visual tests, hand preference tests, eye preference tests and foot preference tests. While some of these tests can be given to groups of children, others must be given individually. The Reading Aptitude Tests by Marion Monroe,¹⁵ for example, are divided into two sections, one of which may be administered to groups of children in about 30 or 40 minutes; the other must be given to each child alone and requires 10 to 15 minutes.

Whether tests of this sort will prove to be much more helpful than mental tests in determining aptitude for reading is somewhat uncertain but there is some evidence which points in that direction. Research indicates that there is much difference in the age at which children acquire the necessary background of experiences, motor development, and mental maturity to make an introduction into the reading processes profitable. These same studies and others have served to point out that meaningful reading at any level is largely dependent upon whether one's background enables one to interpret what one reads. That some of the modern readers ¹⁶ have recognized this is indicated by the fact that before certain units of material are read, pictures are discussed and information built up which will make the material meaningful.

¹⁵ Published by Houghton Mifflin Company.

¹⁶ See, for example, the *Child Development Readers*, published by Houghton Mifflin Company.

Arithmetic

Problems of Measurement in Arithmetic. As we have previously pointed out, measurement in arithmetic does not meet with some of the same difficulties as are encountered in such subjects as reading and language. In the first place, the objectives of teaching arithmetic are somewhat more nearly fixed grade by grade than are those in any other school subject, with the possible exception of spelling. Then too, research workers have been able to break up the processes involved in acquiring facility in the use of numbers in a way which has not yet been done in most of the other subjects. This does not mean that all of the problems of measurement in this field have been solved; neither does it mean that arithmetic is a simple subject. The following classification (by Miss Merton¹⁷) shows how many skills are required and indicates the materials which a comprehensive test in the subtraction of whole numbers must include.

1. The 100 subtraction combinations.
2. Three ideas in one's subtraction concept:
 - Taking away idea: $15 - 7$, 7 and 15.
 - Adding idea: What number added to 7 equals 15?
 - Difference idea: 15 is how many more than 7?
3. The meaning of the following terms: Minus, less, subtrahend, minuend, borrowing, difference, remainder.
4. The meaning of the subtraction sign.
5. That the complete minuend must always be larger than the complete subtrahend.
6. That in writing the example, units must be placed under units, tens under tens, etc.

¹⁷ Merton, E., *Remedial Work in Arithmetic, Second Yearbook of the Department of Elementary School Principals*, pp. 395-411, Washington, D. C., 1923.

7. That one must begin at the right and work to the left.
8. That the order of units in the subtrahend must be subtracted from the same order in the minuend.
9. How to proceed when the first number to be subtracted in the minuend is larger than the corresponding number in the subtrahend.
10. That one must not borrow unless the number in the subtrahend is larger than the corresponding number in the minuend.
11. How to proceed when a number of the subtrahend is larger than the corresponding number of minuend; i.e., borrowing.
12. What it means to place a 1 in front of a number when necessary:

$$\begin{array}{r} 423 \\ -219 \\ \hline \end{array}$$

1 adds in terms of 10

13. What it does to the next number in the minuend when a 1 has been placed before the following number.
14. Must be able to remember the new number made through borrowing

$$\begin{array}{r} 628 \\ -239 \\ \hline \end{array}$$

After subtracting 9 from 18, the child is dealing with 11, not 12.

15. How to proceed when the need for borrowing and no borrowing are met alternately in the example.
16. How to borrow when two or more successive digits in the subtrahend are larger than the corresponding digits in the minuend.
17. How to proceed when there are fewer figures in the subtrahend than in the minuend.
18. How to proceed when the last subtraction takes place with the subtrahend and minuend the same:

$$\begin{array}{r} 649 \\ -623 \\ \hline \end{array}$$

(The zero must not be placed in the remainder.)

19. Ability to handle a zero or a succession of zeros in the subtrahend.
20. Ability to handle a zero or a succession of zeros in the minuend.
21. How to check for correct answers.

When it is remembered that this large list is for a single process for whole numbers only and that there are, no doubt, many skills involved in problem solving also, the complexity of the subject becomes apparent. Every operation, no matter how simple it appears to one who has made the processes automatic, involves many separate steps, each of which must be given attention except in the case of gifted children.

When the skills necessary to the solution of various types of problems have been ascertained, there still remain a number of problems in testing, some of which are also problems of teaching. What problems, for example, should one be able to solve on completion of a given grade or upon completion of the elementary school? Many modern textbooks have been written with social utility as the chief criterion for content. While this basis may be open to some criticism, it seems logical for test makers to follow the lead of the textbook writers. If the ability to deal with abstractions in arithmetic is a desired objective, then test makers must govern themselves accordingly. Problems in cube root, carpeting, paper hanging and the like, having generally disappeared from textbooks, would hardly make suitable problems for an arithmetic test.

When the material of the test has been selected one still needs to decide whether the pupil's score is to be determined by the rate at which he is able to solve the problems, by his accuracy in their solution, by his ability to solve more and more difficult problems, or by his ability to do all three of these. Some writers have indicated

that since there is considerable correlation between speed and accuracy, only one of these need be tested. Such reasoning is fallacious, for it does not recognize the very real possibility that speed and accuracy do not go hand in hand for the individual. The correlation probably exists because the same habits which produce speed also tend to produce accuracy. In other words automatization which makes for rapidity in calculations also makes for greater accuracy. Since the correlation is by no means perfect, there is, however, considerable possibility that a given child may be slow but accurate and another rapid but inaccurate.

There are a number of possible classifications of tests in this field. One might classify them as rate tests and accuracy tests, but so many are found that emphasize both, that such classification would hardly be practicable. The same might be said if one were to classify under the headings of rate tests and power tests or computational and reasoning tests. The writer has, therefore, chosen to classify them according to the purposes which they serve, despite the fact that there, too, we find some overlapping.

Survey Tests. Probably because of the apparent ease of constructing tests in arithmetic there have appeared a considerable number of tests in this subject. Particularly numerous are those of the survey type.

Among the tests more commonly used are probably those appearing in the various test batteries. For the most part they consist of two divisions; one in which the pupils find tests of the fundamental processes and the other in which there are reasoning problems of various sorts.

The Clapp-Young Arithmetic Test ¹⁸ illustrates the concrete problem test often used for survey purposes. Designed for grades 5 to 8, it consists of twenty-five prob-

¹⁸ Published by Houghton Mifflin Company.

lems of the usual reasoning type. Instead of asking the pupils to write their answers, the authors have supplied four answers, only one of which is correct. This one is to be designated by placing a cross in the square before it. In this way it differs from most tests of this type. Space is also provided in which the pupil does his computations so that a teacher who wishes to do so may check over the pupil's work.

Another unique feature is that on the inside pages, which the pupil does not see, is found an explanation of the probable cause when a pupil checks a wrong answer. For example, the first problem in Form A reads as follows: "On a boat there were 319 men, 257 women, and 42 children. How many people were there on the boat altogether?" The suggested answers are 996, 628, 618, and 534. If 996 is given as the answer, the authors indicate that the numbers were arranged by the pupil like this:

319
257; if 628 is the given answer, the pupil made a mistake
42

in adding the second column; whereas, the answer 534 is obtained by adding 319 and 257 and subtracting 42.

Thus the test may serve a very useful purpose in diagnosing pupil difficulties and suggesting remedial treatment for each pupil. When thus used the test is really diagnostic of many of the difficulties encountered in problem solving. Since the test is a bit short, its reliability is not as high as might be desired; hence, for more accurate survey measurement, a reasoning test may be added to advantage.

The Compass Survey Tests ¹⁹ in arithmetic are arranged in different parts, each part dealing with addition, subtraction, multiplication, division, percentage, or general

¹⁹ Published by Scott, Foresman and Company.

problems. They might be called analytical tests since they indicate whether the chief difficulty lies in one or more of these major divisions.

Survey tests in arithmetic may be used to determine if pupils seeking to enter a given grade are prepared to do the arithmetic work of that grade. Again they may be used for grouping children for instruction in arithmetic or for determining whether a given grade has made as much progress as have other pupils of the country who are in the same grade. Finally, some of them may be used for analyzing or even diagnosing difficulties which pupils are experiencing.

Diagnostic Tests. The Buswell-John Diagnostic Chart,²⁰ which was devised after an individual analysis of the work of elementary school pupils, serves as a very helpful device in checking up the work of pupils who are doing unsatisfactory work in arithmetic. The authors recommend that it be used with only one pupil at a time; the pupil to do "all of his thinking aloud" while the teacher notes not only the answers, but particularly the child's method, since it is in the method that the key to difficulties is usually found. No corrections are made at the time, but remedial suggestions are supplied later. The four fundamental processes are all represented on the chart, but only one is to be used at a time. As the pupils work the teacher checks on her copy of the chart (which differs from the pupil's copy by having these suggested methods) the methods employed by the child. Following is the list of habits observed in pupils' work in division.

- d₁ Errors in division combinations
- d₂ Errors in subtraction
- d₃ Errors in multiplication
- d₄ Used remainder larger than divisor

²⁰ Published by Public School Publishing Company.

- d5 Found quotient by trial multiplication
- d6 Neglected to use remainder within problem
- d7 Omitted zero resulting from another digit
- d8 Used wrong operation
- d9 Omitted digit in dividend
- d10 Counted to get quotient
- d11 Repeated part of multiplication table
- d12 Used short division form for long division
- d13 Wrote remainders within problem
- d14 Omitted zero resulting from zero in dividend
- d15 Omitted final remainder
- d16 Used long division form for short division
- d17 Counted in subtracting
- d18 Used too large a product
- d19 Said example backwards
- d20 Used remainder without new dividend figure
- d21 Derived unknown combination from known one
- d22 Had right answer, used wrong one
- d23 Grouped too many digits in dividend
- d24 Error in reading
- d25 Used dividend or divisor as quotient
- d26 Found quotient by adding
- d27 Reversed dividend and divisor
- d28 Used digits of divisor separately
- d29 Wrote all remainders at end of problem
- d30 Misinterpreted table
- d31 Used digit in dividend twice
- d32 Used second digit of divisor to find quotient
- d33 Began dividing at units digit of dividend
- d34 Split dividend
- d35 Used endings to find quotient

Provision is also made for recording other habits or peculiarities as indicated in the following quotation from the Manual of Directions:

As the child works, the teacher should check on the Diagnostic Chart the types of habits which occur, at the same time recording the child's procedure in the space opposite the examples on the teacher's chart. The most satisfactory way to do this is to make a record of the habits observed in the exact words of the pupils, at least for the first few times. If the habit appears later with other examples, it is sufficient to refer back to the earlier procedure. The result of the diagnosis should be that the teacher has a clearer knowledge of the specific habits which are responsible for the pupil's poor work.

Remedial measures are suggested in a monograph published by the University of Chicago Press.

This chart and its method of use serve to emphasize the fact that difficulties are individual matters and need to receive individual attention. It will probably not be necessary for the teacher to study many pupils in this intense manner, but almost every teacher will find one or more pupils who will respond only to treatment of this sort.

*The Compass Diagnostic Tests in Arithmetic*²¹ consist of twenty separate tests, "each covering from one to seven basic arithmetic processes or skills commonly taught in grades two to eight."

The Manual of Directions gives the following suggestions concerning the main situations in which the tests are useful.

1. Test upon completion of some important instructional unit such as long division or addition of fractions in order to determine which pupils are ready to begin a new unit of instruction.
2. Test certain or all of the pupils of a new entering

²¹ Published by Scott, Foresman and Company.

class, particularly when the teacher has no adequate information about the previous preparation of these pupils.

3. Test individual pupils who fail to make normal progress in learning a new skill. Very often the mastery of previously taught topics essential to the new skill has not been secured.
4. Test new pupils for purposes of grade location and classification.
5. Test at intervals entire grades throughout the school system as a check on the general efficiency of arithmetic work in the schools. An economical plan might be to select one test for each grade, varying the tests the second year by choosing a different set of tests, grade by grade.

Following the scoring of the test, the authors recommend the use of the Economy Remedial Exercises in Arithmetic which are so arranged that the teacher will know which to give pupils with certain difficulties disclosed by the tests. Such materials in the hands of a careful teacher should prove very helpful indeed. There may be some danger that the inexperienced teacher, finding such materials prepared for her, will come to the erroneous conclusion that the application of the test and the remedial instruction material will solve all individual difficulties in some mysterious mechanical fashion. That such is not the case will be recognized by anyone who has studied the nature of individual differences. In general, such tests serve to locate approximately where pupil skills are inadequate without showing why. There is also this to be remembered, that the tests, comprehensive as they are, do not test all phases of arithmetic.

The Learning Cycle Diagnostic Tests and Remedial Units. This series of diagnostic tests by Torgerson and

Matthies ²² consists of the following sub-tests, each printed separately so that they may be given as occasion arises:

- The Addition of Whole Numbers
- The Subtraction of Whole Numbers
- The Multiplication of Whole Numbers
- Division of Whole Numbers. One-place Divisors
- Division of Whole Numbers. One- and Two-place Divisors
- Division of Whole Numbers. Two- and Three-place Divisors
- Meaning of Fractions
- The Addition of Fractions
- The Subtraction of Fractions
- The Multiplication of Fractions
- The Division of Fractions
- The Meaning of Decimals
- The Addition and Subtraction of Decimals
- The Multiplication of Decimals
- The Division of Decimals
- The Meaning of Percentage
- The Three Common Uses of Percentage
- Discount
- Interest

Pupils who fail to achieve the desired goals in any one or more of these tests may then be given the proper remedial units which have been carefully prepared in such a way as to give the pupil the particular type of exercise needed to help him overcome his difficulties. Since the material is self-instructive, the pupil may work at his own rate of speed. At the end of this practice a test is provided which enables the teacher to determine whether the pupil has mastered the skills and abilities involved. If he has not, he is directed to additional practice material which is also provided by the authors. While the writer of this volume has not had first-hand experience with these materials, it

²² Published by E. M. Hale and Company, Milwaukee, Wisconsin.

appears that they should be extremely valuable for the teacher of pupils who vary markedly in their arithmetical abilities.

Practice Tests and Exercises. While tests have been used mainly to give the teacher an index of pupil achievement, there are those who feel that most good can be accomplished by tests if they serve more directly an instructional purpose. For this reason there have been developed a considerable number of workbooks and practice tests designed primarily to give the pupils an opportunity to check themselves in their own work and to note their own progress or failings. We have already mentioned the Economy Remedial Exercises.²³ These are arranged on cards which contain rectangular holes underneath each problem affording an opportunity for pupils to write their answers on a sheet placed beneath the card. Each pupil is supplied with a card containing the problems which the diagnosis reveals that he needs, thus enabling the teacher to have the various pupils working at different tasks according to their peculiar needs.

*The Clapp Drill Books in Arithmetic*²⁴ also furnish practice material which supplements that found in the textbook. One drill book is available for each of the grades from 3 to 8 and the material presented is suited to the grade for which it is designed. In addition to the exercises in the fundamental processes there are in each book a series of "concrete problems" of the type needed by pupils of the particular grade. Pupils write their answers on a separate sheet of paper so the books can be used again and again.

Another carefully constructed series of workbooks is known as Teaching and Practice Exercises in Arithmetic

²³ Published by Scott, Foresman and Company.

²⁴ Published by Silver Burdett Company.

by Buswell and John.²⁵ There are separate books for each of the grades from 3 to 6. Pupils work the problems on a sheet, tear out the sheet and check their problems with the answers found on the top of the reverse side of the sheet. At the top of each sheet are also found a few suggestions to help the pupil about to undertake the work.

Other practice exercises designed for other types of work in arithmetic are the Clark-Otis-Hatton Instructional Tests in Arithmetic for Beginners, for grades 1 to 3;²⁶ and the DeMay-McCall Standard Test Lessons in Fractions, for grades 5 to 8.²⁷

Probably material of this sort will be most useful if children work with it only once or twice a week. One of the advantages is the novelty that such material offers, which, if the material is presented too often, is soon exhausted.

Teacher-Made Tests in Arithmetic. In contrast with the situation which exists in reading, testing by means of tests which the teacher herself makes is very common. Very often such tests will concern themselves with the fundamental operations which all pupils are expected to master thoroughly. Unless the teacher takes rather elaborate precautions against it, she is likely to become a victim of the tendency to repeat certain combinations very frequently and to neglect others. An analysis of some of the texts published less than twenty years ago revealed, for example, that certain number combinations were repeated many more times than other combinations. It was also revealed that the combinations on which least practice was pro-

²⁵ Published by Wheeler Publishing Company, Chicago.

²⁶ Published by World Book Company.

²⁷ Published by Bureau of Publications, Teachers College, Columbia University.

As the teacher provides for the combination $6 + 8$ she puts a check in the appropriate square as shown above. Following this procedure she will know exactly what combinations are used when her test is complete. A similar device will be found convenient for noting the combinations on which pupils most often make mistakes. Drill exercises to remedy existing weaknesses can then be more intelligently provided. Similar charts can, of course, be made for subtraction, multiplication, and division.

PROBLEMS

1. Make a list of your own reading activities for a day. Note particularly those which are not connected with your school work.
2. Is it advisable to test elementary school pupils on their ability to read and interpret poetry? Why or why not?
3. Do you favor arranging questions in reading tests in such a way that the pupil reads with the same purpose for a long time, or do you prefer a miscellaneous arrangement of questions in such a way that the pupil must have several objectives in mind when reading the given selection?
4. Discuss the advisability of writing arithmetic textbooks with social utility as the chief criterion for content.
5. Make up a test of twenty-five different combinations in multiplication. Use a chart similar to the one on the previous page to indicate the combinations used.
6. Why has the field of arithmetic been subjected to more careful analysis and measurement than several other fields?

BIBLIOGRAPHY

Reading

- Betts, E. A., *Diagnosis of Reading Disabilities, Teacher's Guide for Use with Telebinocular Equipment*, Keystone View Company, 1934.
- Dearborn, Walter F., Motivation versus "Control" in Remedial Reading, *Education* 59:1-6, September 1938.
- Donnelly, Helen E., The Remedial Reading Classroom, *Education* 59:31-36, September 1938.
- Durrell, Donald D., Basic Abilities in Intermediate Grade Reading, *Education* 59:45-50, September 1938.
- Gates, A. I., *A Reading Vocabulary for the Primary Grades*, Bureau of Publications, Teachers College, Columbia University, 1926.
- Gates, A. I., *The Improvement of Reading (Revised Edition)*, The Macmillan Company, 1934.
- Greene, H. A., and Jorgensen, A. N., *The Use and Interpretation of Elementary School Tests*, Chapter XIII, Longmans, Green and Company, 1936.
- National Society for the Study of Education—Second Report on Reading, *Thirty-Sixth Yearbook*, Part I, 1937.
- Patterson, S. W., *Teaching the Child to Read*, Doubleday, Doran and Company, 1930.
- Sangren, Paul V., The Measurement of Achievement in Silent Reading, *Bulletin of Western State Teachers College*, Kalamazoo, Michigan, 1927.
- Smith, H. L., and Wright, W. W., *Tests and Measurements*, Chapter IX, Silver Burdett Company, 1928.
- Sullivan, Helen Blair, A New Method of Determining Capacity for Reading, *Education* 59:39-45, September 1938.
- Tiegs, E. W., *Tests and Measurements for Teachers*, Chapter XVIII, Houghton Mifflin Company, 1931.

Arithmetic

- Brueckner, L. J., *Diagnostic and Remedial Teaching in Arithmetic*, John C. Winston, 1930.
- Brueckner, L. J., and Melby, E. O., *Diagnostic and Remedial Teaching*, Chapter VII, Houghton Mifflin Company, 1931.
- Greene, H. A., and Jorgensen, A. N., *The Use and Interpretation of Elementary School Tests*, Chapter XII, Longmans, Green and Company, 1936.
- Morton, R. L., *Teaching Arithmetic in the Elementary School, Volume 1, Primary Grades*, Silver Burdett Company, 1938.
- Morton, R. L., *Teaching Arithmetic in the Elementary School, Volume 2, Intermediate Grades*, Silver Burdett Company, 1938.
- National Society for the Study of Education—Research in Arithmetic, *Twenty-Ninth Yearbook*, Part II, 1930.
- Smith, H. L., and Wright, W. W., *Tests and Measurements*, Chapter V, Silver Burdett Company, 1928.
- Tiegs, E. W., *Tests and Measurements for Teachers*, Chapter XVIII, Houghton Mifflin Company, 1931.
- ,

•

**TESTS IN LANGUAGE, HANDWRITING,
AND SPELLING**

•

Chapter VI

TESTS IN LANGUAGE, HANDWRITING, AND SPELLING

Language

THE OBJECTIVES of language instruction may be stated in general terms as the acquisition of ability to use correct, fluent, and interesting English in the life situations in which language is used. Many such occasions arise both in and out of school as is indicated by the following list of general life situations in which language is used: ¹

6

I. Life situations in which spoken English is used

A. Conversations

At the table, at social gatherings, in discussion groups, at public gatherings, in public places, during introductions, during calls, at interviews, in greetings and partings, in asking directions, in telephoning.

B. Meetings

1. Informal proceedings such as classes, auditorium exercises.
2. Formal proceedings of organizations, clubs, committees.

C. Practical discussions

1. Speeches of felicitation, dedication, presentation of gifts, acceptance of gifts, introduc-

¹ *Fourth Yearbook*, Department of Superintendence, pp. 259-260.

tion of speakers, inauguration speeches, speeches upon retiring from service, substitute or impromptu speeches.

2. Reports of meetings, conferences, visits, illustrated lectures, demonstration talks.
3. Persuasive talks as in membership drives, political campaigns, school campaigns for thrift, health, cleanliness; as in applying for the position of office boy or paper carrier; as in selling tickets to school entertainments.
4. Messages and announcements of games, lectures, exhibits, entertainment, meetings.
5. Explanations and directions as to how to make a radio, a cake, or a flower box; how to go to a park or a railway station; how to iron a dress or care for the children.

D. Anecdotes and stories

1. Telling anecdotes and stories to children in the home, school, or social group.
2. Telling anecdotes and stories to adults at social functions, on the train, at the dinner table, at informal gatherings of friends, to people who are sick or in trouble, at public meetings.

II. Life situations in which written English is used

A. Letters

1. Business letters to firms for information, for supplies.
2. Social letters to school friends, to parents, to children in other communities.
3. Informal notes: excuses, invitations.
4. Formal notes.

B. Notices of games, lectures, exhibits, entertainments, meetings.

C. Reports of committee to school or class; of delegate to class or school council; official, president of school council; financial, money saved by class each week; minutes of council or club;

- reviews, books, articles, speeches, plays; of observations or experiments.
- D. Note taking for preparation of papers, stories, discussion, and reports.
 - E. Filling out forms, mail order blanks, application for money orders, checks, deposit slips, test forms, telegrams or cablegrams; information blanks or questionnaires, budgets.
 - F. Making a bibliography.
 - G. Creative writing for papers, clubs, class, newspaper or magazine articles in school or local paper, diaries, imaginative writing, such as stories, poems, plays.

In addition to these situations one should not overlook the very important use of language as a tool of thinking—thinking which may not find outward expression in either written or oral form.

The test expert has met with difficulty in attempting to measure many of the outcomes of language instruction. Particularly is this true of the first group of situations listed above; namely, in the case of spoken English. Because test development has proceeded largely along the lines of group testing, spoken English has been neglected. Considering the very great use for spoken English, this is unfortunate but so far unavoidable. Perhaps methods will be developed for precise measurement in the field of oral English. The work of Dr. H. A. Greene² gives promise in this direction and certainly is of value in the analysis of errors in oral as well as written composition.

Another difficulty confronting the test maker is the fact that definite objectives for each year of work have rarely been set up. As a matter of fact, the matter of analyzing or even identifying the various basic language skills has not progressed very far. Most of the researches which have

² See *Elementary English Review* for March, April, May, 1933.

been made have been concerned with the identification of errors rather than of basic and essential skills. The errors are more easily identified and probably less numerous, though errors are possible in every situation and perhaps do actually occur.

From the teacher's viewpoint, language is a difficult subject also in part because it is a subject in which she rarely if ever initiates the child. Few children come to school with any well-developed number concept or any reading habits, but practically every child has mastered a few hundred or even several thousand words before his school life begins and has learned to manipulate these words in various ways—almost always in some incorrect ways. Further, much of the child's activity in the field of number is, at least while his first rudimentary concepts are being formed, confined to the schoolroom, whereas language activities are at least as numerous outside the schoolroom as they are in school. There are, moreover, many influences which tend to encourage a child to use an incorrect language form while very seldom is any pressure brought upon a child to give an incorrect answer when confronted with a number combination. Hence, there is no particular assurance that a child who has been taught what constitutes correct form will use that form in his written work and more especially in his conversation.

While some language tests measure many phases of language, most of them purport to deal either with language usage, grammar, or composition.

Language Usage Tests. The tests which fall under this caption probably should be called language knowledge tests since what they test is not habitual usage but the child's knowledge of what constitutes the correct form.

Often there is a vast difference. These tests are based mainly upon the most common errors made by children or adults. Many excellent studies by Charters, Clapp, Wilson, and others have revealed what these errors are and have indicated that, while a few of them are peculiar to given localities, most of them are found in every section of the United States.

Among the best tests of this type is the Pribble-McCrory Diagnostic Test in Elementary Language³ which is designed for grades 3 to 6 inclusive. The various parts of the test are concerned with language usage, capitalization, pronunciation, and punctuation. The test has been very carefully constructed to include measures of many of the significant outcomes which pupils should realize by the end of the sixth grade. That lack of agreement as to what constitutes the objectives of language instruction for the various grades makes test construction in this field a bit difficult is indicated in the following quotation from the authors:

The correct use of "broken" is given first presentation in the second grade in some courses of study, in the third grade in others, in the fourth grade in others, and in the sixth grade in another. The same lack of agreement in initial presentation of the possessives of nouns, of pronoun constructions, and in redundancy situations occurs. In the matter of capitalization and of punctuation, except for a few fundamentals (e.g., initial capitalization and final punctuation of a sentence) placement practice is as variable here as in correct usage. For example, capitalization of "South" and other words referring to sections of the country are introduced in the third grade in some courses of study, and even in the sixth grade in others. The use

³ Published by Lyons and Carnahan.

of quotation marks in their grade placement is equally variable.

Probably the most unusual feature of the test is the pronunciation test.

The Clapp-Young English Test ⁴ is somewhat similar to the Pribble-McCrory test and deals with capitalization, punctuation, word form, and grammar. Since the test is printed in the Clapp-Young Self-Marking form, the teacher needs only to open the booklet and count the squares in which crosses appear in order to obtain the pupil's score. Alongside each square is printed a statement of the point or rule embodied in the exercise. Diagnosis is thus facilitated and the teacher needs simply to observe the squares within which marks do not fall in order to determine the points on which her pupils may be given remedial drill and practice. The test is designed for grades 5 to 12.

The Charters Diagnostic Language Test ⁵ consists of four separate tests, one of which is concerned with pronouns, one with verbs, and the other two with common errors of a miscellaneous nature.

The Franseen Diagnostic Tests ⁶ are similar to the Charters tests but are a bit more elementary in form and appear to have been more scientifically prepared, though no information concerning their preparation is supplied with the tests. Part I deals with pronouns, Part II with verbs, and Part III with varied constructions. The tests are designed for grades 3 to 8 and are, in contrast with the Charters test, thoroughly objective.

The Wilson Language Error Test ⁷ is of a somewhat

⁴ Published by Houghton Mifflin Company.

⁵ Published by Public School Publishing Company.

⁶ Published by C. A. Gregory Company, Cincinnati, Ohio.

⁷ Published by the World Book Company.

different form, consisting of a story containing a number of errors which the pupil is to find and to correct.

No time limit is fixed but pupils are instructed to "work as quickly and carefully as possible." The pupil's score is dependent upon the number of errors corrected, no penalty being given for correct forms made incorrect by the pupil. The test is thus essentially a "proof-reading" test and is useful in making pupils conscious of the errors which they are likely to make. There are three stories, each to be used at a different time but in the same manner.

Grammar Tests. Independent grammar tests have not ordinarily been developed for the elementary grades. The reason for this can be found chiefly in the growing realization that a knowledge of grammatical rules contributes little, if anything, to the ability to use language properly. Such few grammatical principles as have been found of probable use are incorporated in some of the so-called language usage tests. It is at least doubtful that all of the children need to be acquainted with grammatical principles in order to use correct speech or to write correctly. On the other hand, when an individual pupil seems to require knowledge of a given principle, there is good reason for giving instruction in the rules of the principle involved. In other words, grammar instruction becomes incidental, just as one often learns the rules of a game as the need arises rather than memorizes them all at once before beginning to play.

How to Use Language Tests. In order to secure the maximum usefulness from language usage tests and indeed from many tests in other fields, it is not sufficient to know that some pupils have difficulties which other pupils do not have; one must know what specific difficulty

each pupil has. A convenient form for indicating this information is the following.

	A. A.	E. A.	J. B.	O. B.	E. C.	F. C.	P. D.	A. E.	J. E.	L. E.	R. E.	S. E.	G. G.	B. H.	G. J.	M. J.	O. J.	P. K.	F. L.	G. M.	M. N.	K. O.	W. P.	R. R.	G. T.	Total
1												x														1
2							x											x								2
3				x					x							x										3
4	x		x	x		x			x			x		x		x			x				x			10
5				x																			x			3
6		x									x	x														3
7				x	x								x								x			x		5
8	x	x	x	x	x	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x		x	x	x	21
9	x						x		x		x				x				x			x				7
10					x						x								x				x			4

The above chart records the errors made by a fifth grade group on the first ten items of the Clapp-Young test. That only one pupil (S. F.) made an error on the first item indicates that all of the other pupils recognize the need for a period at the end of the sentence. This is a matter, then, for which the time of the entire class need not be taken. On the other hand, item number 8, which calls for the apostrophe to indicate possession, needs to be called to the attention of practically all of the pupils, since twenty-one pupils did not detect this error. By the use of such a device attention is called to the difficulties of each pupil and appropriate practice material can be supplied. It must be remembered, however, that not all of the ability called for in most language tests is required in a given grade. If, for example, the correct use of the apostrophe is not called for in the fifth grade, the teacher will pay no attention to errors on this point except where pupils manifest an interest and raise questions about it.

Composition Scales. The language usage tests and grammar tests emphasize solely language form as contrasted with language content. The objectives of language instruction, however, include improvement in both form and content. As an illustration of the difference in these

two important phases Monroe and Streitz⁸ point out that those who are interested primarily in the form objectives would state the language objectives somewhat as follows:

1. Ability to arrange and punctuate the heading of a letter;
2. Ability to choose and write an appropriate form of salutation;
3. Ability to formulate an appropriate first sentence in such situations as, (a) when replying to a letter, (b) when answering an advertisement, (c) when reporting a complaint;
4. Ability to begin each sentence with a capital letter and to place a period after each declarative sentence;
5. Ability to determine when a new paragraph is required.

Content objectives, on the other hand, might be stated in functional terms as,

1. To provide new experiences which stimulate growth in language;
2. To use a wide range of topics which originate in small group discussions;
3. To help children keep adequate records of certain school experiences;
4. To utilize the school assemblies, festivals, exhibits, newspaper and other school activities to provide standards for composition work;
5. To encourage original and creative efforts both in prose and poetry.

Neither of the above lists is intended to be complete or even comprehensive since an exhaustive list, particularly of content objectives, would be voluminous indeed.

As was pointed out above, language usage tests and

⁸ Monroe, Walter S., and Streitz, Ruth. *Directing learning in the Elementary School*. Doubleday, Doran and Company, 1932. P. 231.

grammar tests neglect the content objectives completely; most composition scales, on the other hand, while not neglecting the form objectives, have paid some attention to content as well. Unfortunately the development of composition scales has not received as much attention as the development of tests devoted to the measurement of correct form. This may be due in part to the fact that the desire for objectivity in measurement seems somewhat less attainable in evaluating compositions. Whatever the cause, the only widely used and often cited composition scales are those developed a number of years ago.

The Willing Scale⁹ resembles in some respects the better known Ayres handwriting scale in that specimens of compositions written by pupils have been arranged in order of increasing merit and have been assigned corresponding numerical values. The poorest composition shown on the scale is given a score of 20; the second a score of 30; and so on up to the best composition which is rated 90. In preparing the compositions to be scored, the pupils are asked to write on any one of a number of suggested topics or on a topic of their own choosing. Twenty-five minutes are allowed for the completion of the stories and pupils are encouraged to make their own corrections within the time limit. In rating the compositions by means of the scale, two qualities are recognized: (1) "story value" and (2) "form value."

Letter-writing Scales. Lewis has developed a series of five composition scales, four of which are devoted to letter writing, as follows: (1) order letters, (2) letters of application, (3) social letters (narrative), (4) social letters (expository). The fifth series is devoted to evaluation of narratives on the theme, "One of My Most Interesting Experiences."

⁹ Published by Public School Publishing Company.

The Clark Letter Writing Test ¹⁰ consists of three parts and is designed to test ability to write both business and social letters. Part 1 consists of a series of questions relating to the salutations and complimentary closes used in both business and social letters. Part 2 is a test of the knowledge of names applied to the various parts of letters, and Test 3 tests ability to organize two letters from sentences and parts of sentences. These tests or scales should prove to be of considerable instructional value even if their usefulness in securing accurate ratings may be questioned.

As was pointed out earlier in this discussion, composition scales have peculiar merit in that they do attempt to evaluate the content of the pupils' compositions. Form may also be evaluated in a somewhat more "natural" situation but it is probably true that most of the form values and difficulties can be discovered more satisfactorily by means of such tests as the Pribble-McCrory, Clapp-Young, or other of the so-called language usage tests. Certainly such tests enable one to make a more rapid diagnosis of pupil difficulties.

It has been shown, however, that by the use of these scales teachers are able to rate compositions in such a way as to be in closer agreement with their own previous ratings and those of other teachers.

Language Practice Materials. As has previously been pointed out, drill must be supplied which will fit individual needs since some pupils make errors of a type which never appear in the work of other pupils. There are several sets of standardized practice exercises now on the market, most of them put up in the form of booklets. Among them may be listed the following which are well known.

¹⁰ Published by Public School Publishing Company.

Guiler—Diagnostic Tests in Punctuation and Remedial Exercises, Rand McNally Company

Mullen and Lanz—Exercises and Tests in English, Ginn and Company

Uhl-Hotz—Practice Lessons in English, Ginn and Company

The Pribble-Brezler¹¹ exercises are on a series of cards. Each card has material designed to correct the more common faults and provides practice in using the correct form. Materials of this sort can be very productive of improvement but only if the practice is well motivated. Symonds and Chase¹² have conducted a study which proves rather conclusively the value of tests as motivators.

Literature Tests. Because the course of study varies so greatly with respect to the literary selections read, very few tests of the knowledge of literature have appeared for use in the elementary school. One of the most widely used is the Stanford Achievement Test in Literature. The test is designed to measure the pupil's acquaintance with certain well-known literary selections, but its validity may well be questioned. Obviously, a pupil might be an omnivorous reader and yet not have read the particular passages selected. This is particularly true, of course, of pupils in the lower grades, but applies, though with somewhat less validity, to the upper grades as well. Evidence of the lack of validity is found in the low reliability coefficients in all grades, the lowest being $r = .22$.

Handwriting

Simply stated, the objectives of teaching handwriting are to see that pupils become capable of writing legibly and

¹¹ Published by Lyons and Carnahan.

¹² Symonds and Chase, Practice vs. Motivation, *Journal of Educational Psychology*, Vol. 29, pp. 19-35, January 1929.

with as much speed as is demanded in practice. While speed of writing may be measured quite easily by the simple expedient of asking pupils to write for a specified time and then counting the letters or words written, legibility has presented many difficulties, the solution of which has awaited the development of more elaborate measuring devices. While much remains to be done in this field, as in all departments of educational measurement, the teacher is now able to indicate to the pupil, or is able to direct the pupil, to determine for himself the progress which he has made from time to time. Further, some of the handwriting scales have aided in directing the attention of the teacher and the learner to the particular defects in writing which must be overcome if legibility is to be increased.

While a number of handwriting scales have been developed, most of them may be classified under two headings; namely, the general merit scales typified by the Ayres Scale, and the analytical scales represented by the Freeman Chart for Diagnosing Faults in Handwriting.

The Ayres Scale. As originally developed in 1912, the Ayres Scale was based upon legibility as indicated by the time required to read the material. There were 24 samples, 8 of which were for users of the vertical style; 8 for users of semi-slant type; and 8 for full-slant writers. Since 1917, when the so-called Gettysburg edition of the Ayres scale appeared, this scale has come into very general use. A part of the Gettysburg Address is used as copy; hence, the name. This scale shows but one slant—the semi-slant, which is most commonly employed, and consists of eight specimens graded from 20 to 90 in accordance with their merit. This scale, or another of the same type, should be placed in each schoolroom so conspicuously that the pupil may readily compare his own writing with the various

specimens and thus evaluate his own work and note his own progress. The use of such a scale is also of considerable help to the teacher in assigning meaningful grades, since the pupil who can see how his own writing compares with the specimens of the scale is much more apt to sense the reasonableness of the grade assigned.

In giving a writing test for the purpose of securing samples for rating, the pupils are asked to write as much of the Gettysburg Address as they are able, in two minutes. Specific directions for administering the tests are found on the scale itself. No effort is made to have pupils write rapidly, but they begin by memorizing the first three lines of the Gettysburg Address so that no time will be lost in referring to copy. The copy is placed on the board, however, to be referred to if necessary. The writing samples are scored according to the number of letters written per minute, and for quality. In scoring for quality, the scorer is instructed to "slide each specimen along the scale until a writing of the same quality is found. The number at the top of the scale above this shows the value of the writing to be measured."

Rating legibility in this manner is admittedly not altogether objective. Some variability in the judgment of various scorers will be found but, as Tiegs¹³ has pointed out, the variability is considerably less even among inexperienced users of the scale than among those who rate without a scale. C. T. Gray¹⁴ has also shown that an inexperienced person, grading three or four hundred specimens and using this scale, can reduce the variability to the point where it is practically negligible.

¹³ Tiegs, E. W., *Tests and Measurements for Teachers*, Houghton Mifflin Company, p. 333.

¹⁴ Gray, C. T., The Training of Judgment in the Use of the Ayres Scale for Handwriting, *Journal of Educational Psychology* 6:85-95, 1915.

Naturally the most important characteristic of handwriting is its quality; yet, perfection in quality is not the aim of school instruction. Studies have indicated that for most practical purposes a quality of 60 on the Ayres Scale is sufficient. In view of the acknowledged tendency for handwriting to deteriorate when attention is no longer directed toward improvement, many schools have established as their goal a quality of 70 for the eighth grade. This gives the pupil a pretty definite objective and if he knows that his formal handwriting drills may be eliminated when this objective has been attained and as long as it is maintained, this knowledge often serves as a powerful motivating influence. There is no good reason why a certain standard for each grade should not be decided upon so that the pupil may know what is expected of him.

Freeman's Chart for Diagnosing Faults in Handwriting. As indicated above, some of the handwriting scales are not so much concerned with determining the general merit of writing as they are with indicating the specific qualities which need to be improved. Such a scale has been devised by Freeman. Five qualities are examined for rating on this scale; namely (1) uniformity of slant; (2) uniformity of alignment; (3) quality of line; (4) letter formation; and (5) spacing, both between letters and between words.

The specimen of handwriting to be evaluated is rated on each of these elements separately, an excellent specimen being rated 5; a fair specimen, 3; and a very poor specimen, 1. The intermediate ratings of 2 and 4 may be used if desired, and if a single quality rating is wanted, it may be obtained by combining the five separate ratings.

Scales of this type are excellent instructional devices, since it is much more helpful to the pupil to realize that improvement may be made by correcting a specific defect,

such as the uniformity of slant or the manner in which he makes an "a," than for him merely to understand in a general way that his handwriting needs improvement.

Other Handwriting Scales. While the discussion above has concerned only two penmanship scales, there are a number of others which have excellent features. The Thorndike Scale, the earliest to be produced, is somewhat similar in construction to the Ayres Scale, except that the samples are arranged in an order determined by the opinion of judges. The Minneapolis Self-Corrective Handwriting Charts, the Pressey Chart for Diagnosis of Illegibilities in Handwriting, the Leamer Diagnostic Practice Sentences in Handwriting, the West Chart for Diagnosing Elements in Handwriting, the American Handwriting Scale, the Courtis Standard Practice Tests in Handwriting, the Palmer Scale, the Zaner Scale, and the New York City Penmanship Scale are among those which will be found helpful.

Current Problems of Measurement in Handwriting. The problems of measurement in handwriting are somewhat different from those found in measuring reading, arithmetic, and most of the other elementary school subjects. Handwriting illustrates a sensori-motor type of learning which does not correlate highly with most of the other school subjects. Thorndike¹⁵ found no correlation at all between writing and general scholarship in the case of adults, while Starch¹⁶ found a correlation of only .31 in the case of children. Perhaps no correlation would be found in the case of children were it not for the tendency of pupils who do well in most of their work to exert

¹⁵ Thorndike, E. L., *Handwriting, Teachers College Record*, Vol. II, No. 2, March 1910.

¹⁶ Starch, Daniel, *The Measurement of Efficiency in Handwriting, Journal of Educational Psychology* 6:106-14, February 1915.

greater effort in handwriting also. Society does not condemn poor penmanship as it does, for example, poor spelling. It is further difficult for the teacher to motivate handwriting when the pupil is aware that out of school most persons write very little, while those who do write much usually resort to the use of the typewriter. Another disturbing factor is the recent trend in the teaching of manuscript writing. Very few scales for the measurement of this type of writing have thus far been produced, despite the fact that the movement has been gaining ground rather rapidly.

Spelling

The Problem of Measurement in Spelling. Although there is some dispute and difference of opinion concerning the objectives of spelling instruction, most authorities would agree that they comprise the following:

- A. To develop the ability to spell correctly the words most commonly needed for expression of thought in writing.
- B. To develop the meaning of words to be spelled.
- C. To develop the ability to recognize correct and incorrect spelling of words.
- D. To develop a desire to spell correctly.

As a rule, only the first of these objectives has been measured in spelling tests because of the apparent unwillingness of test makers to depart from the traditional method of testing spelling; namely, by having pupils write the words or sentences dictated by the examiner. A few experiments have been conducted, however, in which other types of tests have been used. Breed¹⁷ believes that the error correction test, that is, a test in which a list of

¹⁷ Breed, F. S., *New-Type Spelling Tests, Elementary English Review*, 7:54-6, March 1930.

words, most of which are misspelled, is presented with instructions to write these words correctly, furnishes a satisfactory method of testing. Breed also experimented with a multiple choice test, as did Cook;¹⁸ Guiler;¹⁹ Pintner, Rinsland, and Zubin;²⁰ and Nelson and Denny.²¹ While there is some conflicting evidence presented in these studies, it appears that the multiple response test is suitable for survey purposes, at least if the following conditions prevail:

1. The test should consist of words which are commonly used in writing vocabularies.
2. The alternate responses should be those misspellings which children make most frequently.
3. The words should preferably be given in context.
4. At least four choices and preferably five should be given.

Such a test has the advantage of being more easily scored, more easily administered, and is not influenced by faulty pronunciation on the part of the examiner or of faulty hearing on the part of the pupil. However, it does not appear to serve well as a diagnostic test since pupils often have to detect different misspellings in this form of test from those which they make when writing the words. It is true, to be sure, that pupils use different spellings also in successive writings of the same word, but probably not to the same extent.

¹⁸ Cook, W. W., *Measurement of General Spelling Ability Involving Controlled Comparisons between Techniques*, *University of Iowa Studies*, 1932, Vol. VI, No. 6, p. 112.

¹⁹ Guiler, W. S., *Validation of Methods of Testing Spelling*, *Journal of Educational Research*, 20:181-9, October 1929.

²⁰ Pintner, R., Rinsland, H., and Zubin, J., *The Evaluation of Self-Administering Spelling Tests*, *Journal of Educational Psychology* 20:108-111, February 1929.

²¹ Nelson, M. J., and Denny, E. C., *The Multiple-Choice Test in Spelling*, *School and Society* 44:15-16, July 4, 1936.

Other forms of self-administering spelling tests may be found to be of value, but at present those mentioned above seem to give the greatest promise.

Spelling Lists. One of the phases of spelling in which considerable advance has been made is in the choice of words which pupils are expected to learn to spell. Texts published fifty to sixty years ago contained such unusual words as "verisimilitude" and "phantasmagoria" and even thirty years ago children were expected to learn to spell "expunge," "gnu," and other words as rarely used by the typical individual in writing. Studies by Ayres, Thorndike, Horn, and others have provided us with lists of words which are most commonly used and it is on the basis of these lists that the modern spellers are constructed. The number of words included in more than ninety-nine per cent of the vocabulary of the ordinary individual in writing appears to be something less than 4,000; therefore relatively few spellers contain more than this number of words. While there is considerable agreement concerning the words to be included, the agreement is by no means unanimous, as has been shown by Selke²² and Frazier.²³ Disagreement as to the proper grade placement is even greater.

Lists from which spelling tests suitable for a given grade may be chosen have been prepared by Ayres and are presented in his spelling scale. As originally constructed, the scale contained 1,000 words but it was further extended by Buckingham to include a total of 1,505 words. The following explanation by the authors will indicate its character:

²² Selke, Erich, A Study of the Vocabulary of Ten Spellers, *Elementary School Journal* 29:767-770, May 1929.

²³ Frazier, C. F., Comparative Analysis of Eleven Elementary School Spellers, *Master's Thesis*, University of Wisconsin, 1931.

All the words in each column are of approximately equal spelling difficulty. The steps in spelling difficulty from each column to the next are approximately equal steps. The numbers at the top indicate about what per cent of correct spellings may be expected among the children of the different grades. For example, if 20 words from column H are given as a spelling test it may be expected that the average score for an entire second grade spelling them will be about 79 per cent. For a third grade it should be about 92 per cent, for a fourth grade about 98 per cent, and for a fifth grade about 100 per cent.

- ✓ While some studies ²⁴ indicate that the words of a given column are not of equal difficulty for children, the scale is of some use in the construction of teacher-made tests, particularly if the teacher is careful to select the words which her pupils have studied.

Another list from which words may be selected for spelling tests is the Iowa Spelling Scales,²⁵ devised by E. J. Ashbaugh. These are arranged in steps of varying difficulty for each grade. For example, in the list for grade VIII, words are listed which average children are able to spell with accuracy of 100 per cent, 99 per cent, 98 per cent, 96 per cent, and so on down to a single word which only 28 per cent of the pupils of this grade can spell. It is thus possible to select a list approximating the difficulty desired. Words from the Horn or Thorndike lists may also be chosen to make sure of including words most commonly used, but no norms are provided.

Spelling Tests. Among the most commonly used spelling tests is the New Stanford Achievement Test, of which there are several alternate forms. The words to be spelled

²⁴ Denny, E. C., Are the Words in a Given Column of the Ayres Spelling Scale of Equal Difficulty for a Given Class? Unpublished study, Iowa State Teachers College.

²⁵ Published by Public School Publishing Company.

are contained in sentences which are dictated by the examiner. The following sentences from Form W indicate their character: >

Stand well back. ‘

The church party is tonight.

Offer your objection to the judge.

The famous gentleman is my cousin.

The occupants are a trifle unusual.

The eminent humorist is my correspondent.

Undoubtedly the schedule is fatiguing.

A poultice (pōl-tis) for grippe is a nuisance.

The words “aqueous” and “anhydrous” are antonyms.

Only the words in bold-faced type are checked for spelling. The pupils are not told, if the directions are followed, that the test is one of spelling. The purpose in not informing them is, presumably, to encourage them to write as they would normally do. The value of this technique is questionable, since some of the pupils will probably believe they are being tested for their ability to capitalize properly; others may think the test is one of penmanship; while others will no doubt believe it to be a spelling test. Hence, the results will not be strictly comparable.

When the number of words which pupils are asked to learn to spell is reduced to about 4,000 of the most common words, it is likely that some pupils will be able to spell them all and a considerable number will be able to spell ninety per cent or more. The optimum difficulty of words for testing, if one is seeking to secure reliability in a test, is about fifty per cent. This fact has led to the inclusion in the Stanford Achievement Tests of such words as “plebiscite,” “anhydrous,” and “seismograph”—words which are admittedly not frequently used, but which hap-

pen to be of about fifty per cent difficulty. Wilson ²⁶ has shown that since the chief purpose in teaching spelling is "to teach the pupil to spell correctly the words which he understands and wants to use often in his written work," the Stanford tests are not wholly valid. In view of the fact that either validity or reliability must probably be sacrificed, it seems that reliability, as the less important, should receive secondary consideration.

In the Modern School Achievement Test battery the spelling test is so arranged that the pupil writes the word to be spelled in a blank space. For example, in the space left in the following sentence, the word "except" is to be written.

"All James went."

In the Unit Scales of Attainment single words are pronounced except in a few instances where two words have the same pronunciation, in which case a short sentence or phrase is given to convey the correct meaning.

In the Metropolitan Achievement Tests the spelling words are also written in column without reference to context.

The Morrison-McCall Spelling Scales ²⁷ consist of eight lists of fifty words each, the lists being of approximately identical difficulty. In administering these the examiner pronounces the word, uses it in a sentence, and then pronounces it again, the pupil writing only the one word. Because of the large number of forms, the tests are particularly useful in certain types of experimental work. Norms are provided for grades 2 to 9 inclusive, and provision is made for conversion of scores into spelling age, T-scores, or grade equivalents.

²⁶ Wilson, Guy M., *The Purpose of a Standardized Test in Spelling*, *Journal of Educational Research*, 20:319-26, December 1929.

²⁷ Published by World Book Company.

A multiple-choice test has been standardized by Nelson and Denny for the Self-Marking Achievement Battery (see Chapter X). Words were carefully chosen from the Iowa Spelling Scales and administered to large numbers of pupils. The types of errors were then tabulated and the four misspellings occurring most frequently were used with the correct spelling to make a five-response test. Each word is given in context so that the pupil is thereby helped in determining the word to be used. The test has good reliability and appears to serve very well as a survey test. The authors do not recommend it for diagnostic purposes, although it may prove to be useful in that capacity also. Norms are provided for grades 3 to 9 inclusive. The pupil indicates the number of the spelling which he believes to be correct and the scoring is greatly facilitated by the self-marking device.

PROBLEMS

1. What are some of the reasons that pupils who are aware of the correct language to use often use incorrect forms?
2. In which situations do persons whose formal education is limited to the completion of the elementary or high school most often use written English? What does this suggest concerning the emphasis in written work?
3. What do you think of the practice of administering detailed objective tests covering such literary works as *Snow-bound*, *Black Beauty*, *Robinson Crusoe*, *Little Women*, etc.?
4. Discuss the following statement: "Little time should be used in the schoolroom for the development of writing habits. If the future citizen does much writing he will either use a typewriter or will dictate to a stenographer who will take her notes in shorthand."
5. Discuss the relative merits of print-like script and cursive writing. Should the pupil who in the early grades is taught print-like script change to cursive writing? If so, at what stage?

6. What, in your estimation, constitutes a proper criterion for the grade placement of spelling words?
7. Some teachers supplement the spelling lists of a published speller with words chosen from those which pupils find in their history, literature, geography, arithmetic, or science work. Do you approve of this procedure?

BIBLIOGRAPHY

Language

- Cole, Luella, *Psychology of the Elementary School Subjects*, Chapter VI, Farrar and Rinehart, 1934.
- Greene, H. A., Research in Elementary Language, *Elementary English Review*, March, April, May, and June numbers, 1933.
- Lyman, R. L., Summary of Investigations Relating to Grammar, Language, and Composition, *Supplementary Educational Monographs*, University of Chicago Press, January 1929.
- McKee, Paul, *Language in the Elementary School*, Houghton Mifflin Company, 1934.
- National Education Association, *Fourth Yearbook of the Department of Superintendence*, 1926, pp. 259-260.
- Smith, H. L., and Wright, W. W., *Tests and Measurements*, Chapter VII, Silver Burdett Company, 1928.
- Tiegs, E. W., *Tests and Measurements for Teachers*, pp. 345-350, Houghton Mifflin Company, 1931.

Handwriting

- Brueckner, L. J., and Melby, E. O., *Diagnostic and Remedial Teaching*, Chapter XI, Houghton Mifflin Company, 1931.
- Freeman, F. N., and Dougherty, M. L., *How to Teach Handwriting*, Houghton Mifflin Company, 1923.
- Gray, C. T., The Training of Judgment in the Use of the Ayres Scale for Handwriting, *Journal of Educational Psychology* 6:85-95, February 1915.
- Greene, H. A., and Jorgensen, A. N., *The Use and Interpretation of Elementary School Tests*, Chapter XVI, Longmans, Green and Company, 1936.
- Starch, Daniel, The Measurement of Efficiency in Handwriting, *Journal of Educational Psychology* 6:106-114, February 1915.

- Stone, Clarence R., *Supervision of the Elementary School*, Chapter XI, Houghton Mifflin Company, 1929.
- Thorndike, E. L., Handwriting, *Teachers College Record* 11: 83-175, March 1910.
- Tiegs, E. W., *Tests and Measurements for Teachers*, pp. 329-337, Houghton Mifflin Company, 1931.
- West, P. V., *Changing Practices in the Teaching of Handwriting*, Public School Publishing Company, 1927.

Spelling

- Breed, F. S., *How to Teach Spelling*, F. A. Owen Publishing Company, 1930.
- Breed, F. S., New-Type Spelling Tests, *Elementary English Review* 7:54-6, March 1930.
- Cook, W. W., Measurement of General Spelling Ability Involving Controlled Comparisons between Techniques, *University of Iowa Studies*, Vol. VI, No. 6, 1932.
- Horn, Ernest, A Basic Writing Vocabulary, *University of Iowa Monographs in Education*, First Series, No. 4, April 1926.
- McKee, Paul, *Language in the Elementary School*, Houghton Mifflin Company, 1934.
- Nelson, M. J., and Denny, E. C., The Multiple-Choice Test in Spelling, *School and Society* 44:15-16, July 4, 1936.
- Selke, Erick, A Study of the Vocabulary of Ten Spellers, *Elementary School Journal* 29:767-770, May 1929.
- Thorndike, E. L., *The Teacher's Word Book*, Teachers College, Columbia University, 1931.
- Witty, Paul A., Diagnosis and Remedial Treatment of Poor Spellers, *Journal of Educational Research* 13:39-44, January 1926.

•

TESTS IN THE SOCIAL STUDIES

•

Chapter VII

TESTS IN THE SOCIAL STUDIES

WHILE THE OBJECTIVES of teaching the social studies in the elementary school have been elaborately stated by some writers and committees, the major ones for the elementary school may be simply stated as follows:

- A. The development of desirable attitudes and ideals
- B. The development of a sympathetic understanding of the problems of society in various parts of the world
- C. The development of power of reasoning along civic, historical, and geographic lines
- D. The acquisition of information

The above order might represent the order of importance attached to the various objectives if one were to ask teachers to evaluate them. However, in chronological sequence they would need to be reversed, since one could scarcely do effective reasoning without first having acquired some information to use in arriving at conclusions, and since attitudes and ideals, assuming that they could be formed, would be very unstable unless based upon a fund of information.

Problems of Measurement

The measurement of attainment of all of these objectives represents real difficulties to the testing expert. Attitudes

of all kinds have not yet been tested as thoroughly as has information, and even in a limited field no method has been found to reveal the pupil's exact attitude. Where attempts have been made at measurement of this sort, one has been forced to be content with what the pupil has designated as his attitude. Frequently the pupil has then designated what he considered the attitudes acceptable to the teacher or test-maker, without revealing whether or not these represent his own real convictions. The Hill Test in Civic Attitudes¹ represents an attempt at such measurement. It consists of twenty items dealing with a considerable range of material as indicated below.

1. In using public property, the good citizen should:
 - a. handle it carelessly because he does not own it.
 - b. take as good care of it as if it were his own.
 - c. use it so as to get the greatest amount of fun and enjoyment out of it.
 - d. take better care of it than if he owned it because it belongs to others.
5. While walking in the park, you notice a boy lying on the street; he has apparently been seriously injured by an automobile. In this case:
 - a. look the other way and pretend not to see him.
 - b. stop a passing auto to take him at once to a hospital.
 - c. get him a drink of water and brush the dust off his clothes.
 - d. try to learn his name and telephone to his friends.
10. An ideal home is one in which:
 - a. the family have a deep affection and consideration for one another.
 - b. the members are a father, a mother, and three children.

¹ Published by Public School Publishing Company.

- c. there is an abundance of good things to eat and drink and wear.
 - d. the furniture is beautiful; books are numerous; and servants do the work.
15. The highest type of courtesy is:
- a. to say or do nothing which will make another person feel uneasy or uncomfortable.
 - b. to be considerate and thoughtful to everyone and to help those who are in trouble.
 - c. to say only such things to others as will make them feel good and think well of you.
 - d. to be polite to all acquaintances when at school, but to speak only to your friends when away from school.
19. You are buying a tennis racket, the price of which is \$7.50. You hand the clerk a ten dollar bill and he gives \$4.50 in change. In this case you should:
- a. keep the two extra dollars and say nothing about the mistake.
 - b. debate with yourself whether or not to return the dollars.
 - c. promptly tell the clerk about his mistake and return the two dollars.
 - d. keep the extra change and return it later because your conscience hurts you.

A pupil may be quite certain that answer c to the 19th question is the acceptable response, and yet be unwilling to act accordingly when such a situation actually arises.

The measurement of thought or reasoning is about as unsatisfactorily done at the present time, in part because of the difficulty of discerning what constitutes reasoning as contrasted with information. The following illustrations from the Van Wagenen American History Scales² (Thought Scale R) will serve to emphasize this point.

² Published by Bureau of Publications, Teachers College, New York City.

3. (67) The Northmen probably came to America as early as the year 1000, nearly 500 years before Columbus and the Cabots sailed from Europe. There is no record of any one else having come to America before the year 1000.

By whom do you think America was first discovered?

While it is possible that a pupil might reason out from the material given a correct response to this item, it is also quite possible that he may have learned that America was first discovered by the Norsemen and in making this response exercises no more reasoning than is used in answering any fact question. The following item presents a similar situation in that the order of our major wars may previously have been memorized.

8. (73) The battle of Lundy's Lane was fought in the War of 1812; the battle of Petersburg was fought in the Civil War; the battle of Monmouth was fought in the Revolutionary War. Arrange the three battles in the order in which they were fought.

1.
2.
3.

It thus appears that what may be a reasoning or thought question for one pupil may be a fact question for another. Just what is being measured is, therefore, unknown to the examiner. This same difficulty presents itself to a lesser degree in the avowed information tests where certain items may be reasoned out by some of the pupils.

The measurement of acquired information in history and civics would, however, be fairly easy and satisfactory if there were agreement as to what constitutes the desirable outcomes in this field. Almost everyone will agree that there is a body of knowledge with which everyone should be familiar, but very little agreement can be

reached concerning the definite items contained in such a body of knowledge.

Writers of textbooks do not agree, in their treatment of historical topics, with the opinions expressed by those who formulate objectives. Horn,⁸ for example, examined the assertion that "the chief purpose of teaching history in the elementary school is to make pupils more intelligent with respect to the more crucial activities, conditions, and problems of present day life." Without going into detail concerning his study one may call attention to the following findings.

1. While the Committee of Eight, on the basis of whose recommendations practically all courses of study were planned, recommended about equal emphasis on political, military, and social or economic phases of history, the modern textbooks gave about 82% of their space to political and military events.

2. Books dealing with modern problems devote less than one-fourth of their historical references to political and military affairs and more than three-fourths to social and economic history.

3. While elementary texts devote only one-fourth of their references to the more recent period since 1861, books dealing with modern problems devote about 85% of their historical references to this same period.

Tests have quite naturally followed the lead of the texts and contain in general materials with the same emphasis as indicated in the above study.

A controversy has also arisen concerning the advisability of teaching history, geography, and civics in a unified and integrated course under the heading of "social studies." Proponents of this procedure argue that the content

⁸ Horn, Ernest, Possible Defects in the Present Content of American History as Taught in the Schools, *Sixteenth Yearbook of the National Society for the Study of Education*, Part I, pp. 156-172.

of these three subjects justifies unification; to quote from one of the chief advocates of unification, Dr. Harold O. Rugg: ⁴ "We have found abundant arguments for the need of a unified, continuous social science curriculum from our study of the way in which materials from the different subjects are demanded for a really successful lesson. How, for example, can one teach children the history of transportation, the history of our westward movement, of the settling of our country, and the exploitation of its great natural resources without constantly calling up facts of topography, location, and the like, which for several generations have been called 'geography'? It can't be done. Furthermore, good teachers do not try to do it. Skilled teachers have always taken their material from wherever it happened to be found, irrespective of how it was catalogued or pigeonholed. How can a pupil obtain a really clear notion of the way thirty-three million immigrants came to this country between 1820 and 1920 without having the facts of the economics of Ireland in earlier decades, of the political relations of Ireland and England, of the economic and political history of Germany, foreign living conditions, the status of agriculture and industry in both northern and southeastern Europe? But the accounts which are necessary to present any of these matters make use of history—political, economic, and social—of geography, of international relations. And clear handling of such important topics forces the teacher to incorporate also materials from economics, political science, sociology."

Adherents of the traditional separation of these subjects argue, on the other hand, that there is such a wealth of material in each field that it is undesirable to combine

⁴ Rugg, Harold O., *Do the Social Studies Prepare Pupils Adequately for Life Activities?* *Twenty-Second Yearbook of the National Society for the Study of Education*, Part II, pp. 1-27.

them except possibly in the primary grades. As a result of this controversy, there has come to be an even greater divergence in the methods and materials of instruction. Whether a unified program eventually results from this agitation or not, it has done much good in calling attention to the need for interrelating the three subjects rather than keeping the subject matter of each in a water-tight compartment.

History Tests

Partly because of the confusion referred to above and partly because of the fact that formal instruction in history has often been postponed until the seventh grade, there are relatively few standardized tests designed for grades below the seventh. Even those which are designed for lower grades cover material which often is not considered until the junior high school is reached.

Among the tests in fairly common use are the Pressey-Richards Tests of Understanding of American History.⁵ These tests are arranged in four parts testing character judgment, historical vocabulary, knowledge of chronological sequence of events, and understanding of cause and effect relationships. Two other tests in American history are the Public School Achievement Tests in History⁶ and the Denny-Nelson Tests in American History⁶ which are divided into two parts, one dealing with the colonial period and the other with the national period. There are a number of other tests available including those found in the various achievement test batteries, but on the whole standardized testing in history has not assumed the proportions of this type of testing in such subjects as reading.

Since in many schools the course of study calls for a con-

⁵ Published by Public School Publishing Company.

⁶ Published by World Book Company.

sideration of the European background of American history, many teachers will find the Renfrow Sixth Grade History Tests ⁷ to be of value. While the tests are purely informational, the author believes that "successful completion of the tests requires a great deal of reasoning as well as knowledge."

Civics Tests

In part owing to the lack of uniform organization of the subject matter and in part owing to the fact that formal civics instruction is left to the junior and senior high schools, the number of tests usable in the elementary school in this field is very small. Among them we may mention the Burton Civics Test ⁸ dealing with political, economic, and social items and the Hill Civics Tests ⁹ which aim to test attitudes, information, and "civic action." The Upton-Chassell Citizenship Scales ¹⁰ are not tests in the usual sense, but consist of rating scales to be used by the teacher.

Geography Tests

As in the case of history and civics, the objectives in geography have in recent years been modified to emphasize to a lesser degree the acquisition of fact and to a greater degree a sympathetic understanding of social conditions and the peoples of the earth. The reasons for this new emphasis have been summed up by J. Russell Smith ¹¹ for the

⁷ Published by O. W. Renfrow, Cincinnati, Ohio.

⁸ Published by the World Book Company.

⁹ Published by Public School Publishing Company.

¹⁰ *Teachers College Record* 23:71-79, January 1922.

¹¹ *Thirty-Second Yearbook of the National Society for the Study of Education*, pp. 33-34.

National Society for the Study of Education in the following words:

The Machine Age and the New World of Closer
Relations

I suspect that few of us fully grasp the vast making-over of our thinking needed before we shall be able to bring our national and international affairs within a reasonable distance of the possibilities of comfort and good living that the machines even now make possible. From Nebuchadnezzar to George Washington our world made its living by the use of human and brute muscles and depended upon the neighboring fields for its sustenance. It carried its freight in wagons and sailboats, wrote with a pen, sent hurried messages on horseback, and heard the news a month or a year late, if it heard it at all. Suddenly the age of machinery has arrived. Our meals, our clothes, and a hundred materials all about us come from the ends of the world. World trade forces itself into the economic side of our lives at every hour of the day, and our geography lessons are properly filled with it. World information is here in every morning paper.

World investment has also arrived with its billions in foreign lands. In many senses we have already become citizens of the world. We did not plan it, nor can we escape the host of unexpected and most perplexing problems that this age of machinery has dumped upon us. Some of these problems have been solved. You stick a stamp upon a letter and it goes safely to any one of several score of far countries. It may even pass safely through war. That is one international problem that has been fairly well solved, but it took a lot of work and planning to get it started, and to keep it going there is continuous cooperation of many men of many nations, every moment of every day and every night.

The travelers' check, the international travel ticket, the international bill of lading are other smooth-running examples of international cooperation.

More than two hundred and fifty different things are already being done by international cooperation, and their number is steadily increasing. But the big problems, the ones that promote international frictions and international jealousies are not yet settled. Many of the best minds in the world are deeply concerned. Can we develop international cooperation fast enough to outrun the forces that make for international war with its almost undreamed of tools of destruction?

There are a hundred different ways to work for better world relations, but they all need to start from a basis of knowledge—knowledge of foreign countries, of foreign peoples, and the conditions and problems that make them as they are.

This is the great opportunity of the geography teacher—to introduce the children of this generation to the country in which they live, and also to the countries and peoples with whom it seems inevitable that they must have ever-increasing contact.

Zoe Thralls,¹² a member of the Society's Committee on the Teaching of Geography, outlines the objectives of geography under three headings, major objectives, concomitant objectives, and ultimate objectives, as follows:

1. Major Objective

The major objective of geographic instruction is to assist in the development of the child through giving him a knowledge of the interrelationships existing between man and his natural environment in specific regions and an ability to apply such knowledge in solving the problems of living. This implies that the child should learn (1) to distinguish between human and natural elements mentioned in reading matter or indicated in landscapes, pictures, models, maps, and graphs and (2) to see in what ways the natural ele-

¹² *Thirty-Second Yearbook of the National Society for the Study of Education*, pp. 201-203.

ments in any given region help to explain the cultural elements that are characteristic of the region.

2. The Concomitant Objectives

The term "concomitant" is used to denote the objectives that are to be reached in the course of attaining the major one. In other words, attaining these objectives is inherent in attaining the major objective, and requisite to it. If geographic instruction is to reach its major goal, it should be designed to assist the child to gain:

1. Concrete concepts, facts, and relationship ideas necessary for the understanding of the characteristic adjustments man has made, is attempting to make, or might make, to the natural environment in any region studied. Many of these facts have to do with the nature and location or the distribution of the natural and cultural features that are *significant* in the understanding of geographic relationships.
2. The ability to secure knowledge of such facts through the interpretation of pictures, maps, globes, words, specimens, models, graphs, textual materials and through the observations of landscapes in one's home locality and in other regions in which one travels. This involves a knowledge of sources of such information and ability to distinguish between facts of much or of little value in geographic thinking.

3. Ultimate Objectives

The term "ultimate objective" is used to designate objectives reached through, or growing out of, the attainment of the major objective. If the major objective is reached, the ultimate objectives gained will be:

1. A knowledge of geographic facts, concepts, and relationships that will enable the individual to give more intelligent consideration to current problems—individual, community, national, and international.

2. An understanding of how the varied problems of peoples are related to differences in natural environment; and, developed through this understanding, an interest in, and an open-minded attitude toward, the problems, achievements, and possible future developments of other peoples.
3. A growing power to sense and grasp the economic and cultural inter-dependence of regions and peoples.
4. A better understanding of the value of natural resources and the need for intelligent use of them.
5. The ability to make a worthwhile use of leisure time through the vitalization of local field trips, of more distant travel, and of reading because of an understanding of the interrelationships between man's working, playing, living, and the elements of the natural environment.
6. The recognition and appreciation of the variety of human labor in the major types of regions through the world, arising from an understanding of man's adjustments to his natural environment.

Since only a few geography tests have been standardized in recent years, many of the objectives, particularly the social objectives, have not been measured in available tests. Most of them have been concerned with information, particularly place-names.

Among the better known tests may be mentioned the Buckingham-Stevenson Place Geography Tests,¹³ the Gregory-Hagerty Geography Tests,¹⁴ and the Wiedefeld-Walther Geography Test.¹⁵ This last test is one of the best and most comprehensive, testing for ability to read geographic materials, for ability to see relationships, for ability to

¹³ Published by Public School Publishing Company.

¹⁴ Published by C. A. Gregory, Cincinnati, Ohio.

¹⁵ Published by World Book Company.

read maps and graphs, for a knowledge of geographic vocabulary, for what is called, for want of a better name, "organization," and for ability to locate various places.

Combination Tests in the Social Studies. Except for those appearing in some of the test batteries, tests designed to measure achievement in the combined social studies courses have been slow to develop. When one considers the recency of development of unified social studies curricula this is hardly surprising, especially in view of the fact that standard testing in this field has never enjoyed great popularity.

Teacher-Made Tests in the Social Studies. In view of the relatively unsatisfactory standardized tests in the social studies teachers should be particularly alert to the possibilities of measuring the results of instruction by well-constructed and carefully devised tests of their own making. Pupils are particularly well motivated by means of objective type tests given as pre-tests and again as end-tests. The simpler factual information and fact-finding abilities should be stressed in the lower grades with gradually increasing emphasis on the more complex abilities in the upper elementary and secondary grades. Tests are needed which cover a relatively small unit of work and which measure the somewhat less tangible outcomes of instruction whether the social studies are taught as separate subjects or as a unified whole.

The following statement by J. Russell Smith¹⁶ admirably sums up the opportunity and responsibility not only of the geography teacher but of all teachers of the social studies.

We teachers of geography know that the names of capes and mountains will fade from the student's

¹⁶ *Thirty-Second Yearbook of the National Society for the Study of Education*, pp. 38, 39.

mind, that many of the rivers and capitals will melt into an indistinct haze, that many, perhaps most of the facts will be gone from our students when, at thirty-five or fifty-five years of age, they turn their votes into the ballot box that decides some world crisis. We, the teachers of geography, should realize that the frequently recurring opportunities of the geography class mean this: that to us more than to all other social agencies combined is given the power to decide whether the future act of the voter shall be an act of respect or disrespect, of sympathy or antagonism, of understanding or ignorant prejudice—whether war shall wreck us all or whether we shall put it into the limbo where now the personal duel resides, buried by a better method. Now that a better way is established, the gentleman finds that he can get along perfectly well without puncturing his fellow man with a rapier or a bullet. This world is so rich, so very rich, in resources and in the scientific, mechanical, and economic possibilities of better living and of a better civilization. These possibilities can only be realized by the working together of large groups of people within national boundaries and across national boundaries. This requires vision, imagination, ambition, and the wide-reaching concepts that can arise from geography well taught.

This opportunity of the geography teacher is made even greater than it seems by the fact that most adult activities are bent toward the realization of desires conceived before the age of fifteen years.

Uses of Social Science Tests. A number of writers have asserted that existing tests in the social sciences not only are defective, but are positively dangerous. They cite as reason for this position the fact that such tests are concerned with information only, whereas the chief objectives are not the mere acquisition of knowledge. However, since the acquisition of information is an important objective—important not only for the sake of storing information but

also to develop reasoning ability and to lead to proper attitudes—the writer feels that this view is somewhat exaggerated. It is true, of course, that if teachers use the content of standardized tests as representing all of the objectives of history, geography, or civics instruction, then harm may result. Most teachers do realize that such tests do take into consideration only a portion of the work which they are attempting to cover and there appears to be no good reason for refraining from testing those things which are now measured, simply because the entire field is not covered.

So far as diagnosis of individual difficulties is concerned, present tests are quite inadequate. Much work needs to be done along this line.

In conclusion it may be said that tests in the social studies are by no means as well developed as those in arithmetic, reading, or even language. Considering the difficulties involved and the comparative youth of the testing movement, this situation can scarcely cause much astonishment.

PROBLEMS

1. What position do you take with reference to the teaching of the social sciences as a unified subject? Justify your position.
2. What do you consider as the chief objectives of the study of civics? What difficulties are there in measuring the attainment of these objectives?
3. Do you think that recent elementary school textbooks in the social studies have improved? If so, in what respect?
4. Do you agree with the statement that existing standardized tests in the social studies tend to maintain the *status quo* of classroom work in these fields? Why or why not?

BIBLIOGRAPHY

- Ayer, Adelaide, Some Difficulties in Elementary School History, *Columbia University Contributions to Education*, No. 212, Teachers College, Columbia University, 1926.
- Brueckner, L. J., and Melby, E. O., *Diagnostic and Remedial Teaching*, Chapter XII, Houghton Mifflin Company, 1931.
- Cajori, M. H., A Social Studies Program for Nine-Year-Olds, *Education* 53:563-66, May 1933.
- Cole, Luella, *Psychology of the Elementary School Subjects*, Chapter III, Farrar and Rinehart, 1934.
- Lindquist, E. F., and Anderson, H. R., Achievement Tests in the Social Studies, *Educational Record* 14:198-256, April 1933.
- National Society for the Study of Education, Second Report of the Committee on Minimum Essentials in the Elementary School Subjects, *Sixteenth Yearbook*, Part I, 1917.
- National Society for the Study of Education, The Social Studies in the Elementary and Secondary School, *Twenty-Second Yearbook*, Part II, 1923.
- National Society for the Study of Education, The Teaching of Geography, *Thirty-Second Yearbook*, 1933.
- Pressey, L. C., Fundamental Vocabulary in Elementary School Geography, *Journal of Geography* 32:78-81, February 1933.
- Pressey, S. L. and L. C., The Determination of a Minimal Vocabulary in American History, *Educational Method* 12:205-211, January 1933.
- Smith, H. L. and Wright, W. W., *Tests and Measurements*, Chapters 10 and 11, Silver Burdett Company, 1928.

•

TESTS IN MUSIC AND ART

•

Chapter VIII

TESTS IN MUSIC AND ART

Music

Objectives of Music Teaching. While some pupils will take up the teaching of music or the production of music as their life work, public school music instruction is not to be considered chiefly in the light of a vocational subject. Rather, the aims are to be thought of mainly in the light of emotional development and the enjoyment of leisure time. From this point of view, the following objectives for the sixth grade cited by Webb and Shotwell¹ are representative of present trends.

1. Every child shall have acquired the use of his singing voice and pleasure in song as a means of expression.

2. Every child shall have learned to enjoy music as something heard as well as something expressed.

3. Every child shall have acquired a repertory of songs which may be carried into the home and social life, including "America" and "The Star Spangled Banner."

4. Every child shall have developed aural power to know by sound that which he knows by sight and vice versa. Every child shall have acquired the ability to sing at sight, using words, a unison song of hymn-

¹ Webb and Shotwell, *Standard Tests in the Elementary School*, pp. 439-440.

tune grade; or using syllables, a two-part song of hymn-tune grade, and the easiest three-part songs; these to be in any key; to include any of the measures and rhythms in ordinary use; to contain any accidental signs and tones easily introduced; and in general to be of the grade of difficulty of folk-songs such as the "Minstrel Boy"; also knowledge of the major and minor keys and their signatures.

5. Every child talented in musical performance shall have had opportunity for its cultivation.

6. The children shall have developed a love for the beautiful in music and taste in choosing their songs and the music to which they listen for the enjoyment and pleasure which only good music can give.

7. The children shall have acquired the ability to appreciate the charm of design in songs sung; to give an account of the salient features of structure in a standard composition after a few hearings of it; to identify at least the three-part song form from hearing; and to recognize and give titles and composers of a reasonable number of standard vocal and instrumental compositions.

8. Above all, the children shall have arrived at the conception of music as a beautiful and fine essential in the well-rounded, normal life.

Types of Standard Tests in Music. While a considerable number of music tests have appeared, the number is not nearly so large as in the case of most of the elementary school subjects. It appears also that the development of music tests has not reached the same level of perfection which characterizes tests in many subjects, as imperfect as such tests often seem to be. The reasons for this situation are numerous but three stand out pre-eminently. In the first place, some phases of music do not lend themselves readily to testing. To discover whether a pupil recognizes a given key signature, the outstanding contributions of a given composer, and mechanical aspects of performance

on a given instrument is not particularly difficult. On the other hand, the measurement of appreciation, which is a major outcome, is very difficult, in part because appreciation itself appears to defy satisfactory definition. Another deterrent to the development of good music tests is the lack of interest and skill in test construction on the part of most students and teachers of music. Finally, and equally apparent, is the lack of training in the various phases of public school music which handicaps those who are familiar with statistical techniques and the mechanical aspects of test construction.

The existing music tests may be classed under either two or three headings depending upon one's preference. We shall classify them as (1) prognosis tests, or (2) achievement tests, and (3) tests of appreciation. Others might prefer to drop the third classification and include such tests in the achievement test group because of the intellectual element which certainly plays no inconsequential part in appreciation.

Prognosis Tests or Tests of Musical Talent. The earliest and best known test of musical talent was developed by Seashore in 1915. The series now used consists of five parts, each of which is found on a double-disk record. The five tests are named below and a brief description of each is added.

1. Sense of Pitch. One hundred pairs of tones are sounded and the examinee is asked to indicate whether the second tone of each pair is higher or lower than the first. The range of difference between tones ranges from 30 vibrations per second to $\frac{1}{2}$ vibration per second.

2. Sense of Intensity. This test also consists of 100 pairs of tones but instead of indicating relative pitch, the pupil is asked to indicate whether the second tone

of each pair is louder or softer than the first. The test embraces a wide range of intensity differences.

3. Sense of Time. In this test, two time intervals are marked off by the sounding of three clicks. "The problem is to tell whether the second time interval is longer or shorter than the first. The standard interval is 1.00 sec., and the varied interval is, in turn, 1.02, 1.05, 1.09, 1.14, and 1.20 sec." There are 100 trials.

4. Sense of Consonance. In this test pupils must judge whether the second pair of tones heard is better or worse than the first pair of tones. Fifty pairs are given and the bases for judgment are smoothness, blending, and purity.

5. Tonal Memory. Fifty trials are given here, representing five degrees of difficulty, the degrees of difficulty being increased by having two-tone patterns in the first 10 pairs; three in the second series of 10; and so on until the last 10 trials are made up of six-tone patterns. Each tonal pattern is repeated with one tone changed. The pupil's task is to identify the changed tone and to indicate whether it was the first, second, third, fourth, fifth, or sixth.

A sixth test on the Sense of Rhythm was included, but appears to have been dropped from the present series.

If the Seashore tests fulfill their function, it should be possible to use them for the following purposes:

1. To classify beginning pupils in music for purposes of instruction.
2. To select those pupils who will probably not profit from much music instruction.
3. To select those pupils who give promise of becoming very proficient in music if given the proper training.

Whether the tests do enable one to select and classify students in such a way appears a bit doubtful. Various studies of their prognostic value arrive at different con-

clusions. It appears quite likely that students who make very low scores on all or most of the tests should not be encouraged to plan an extensive musical career. High scores, on the other hand, do not guarantee that the person making them will be an outstanding success, for after all, performance is dependent upon a number of other factors. Due credit must be accorded Seashore for an outstanding piece of pioneer work in a field which certainly is not one of the easiest in which to work out tests of this sort. It is quite likely that by building on this early work someone will develop a more helpful instrument of prognosis in the field of music.

A series of tests of a somewhat similar nature to the series devised by Seashore is the Kwalwasser-Dykema or K-D Music Tests.² While they attempt to measure musical capacity, they do also make some use of the pupils' previous training in the field. The series consists of ten tests, each shorter than those of the Seashore tests and are thus recorded on only five double-disk records. Instead of using both sides for a single test, one test appears on each side of the five records. The following brief description will indicate the character of each test.

1. **Tonal Memory.** This is measured by means of 25 pairs of music patterns varying from patterns of four notes to those containing nine notes. Pupils are asked to listen to the two pairs and then to indicate whether they contain the same or different tones.

2. **Quality Discrimination.** In this test there are 30 pairs of motives and the problem for the pupil to decide is whether the two motives of a pair were played by the same or different instruments. The instruments used vary, the first item, for example, using the clarinet and trombone.

3. **Intensity Discrimination.** This test is similar to

² Published by Carl Fischer, Inc.

the Seashore test of intensity except that there are only 30 pairs and instead of producing a single note in every case, simple chords are used for fifteen of the pairs. There is also this difference, that while in the Seashore test the intensity always varies, there are some cases in the K-D test where the intensity in the two items of a pair is identical.

4. Tonal Memory. This test is somewhat more complex than the first test, since the pupil must here decide whether the completion of the tonal pattern is up or down. There are 30 tonal patterns of four notes each. A fifth note which would complete the pattern is not sounded and the pupil is asked to indicate whether this fifth note should be higher or lower than the fourth note in the pattern.

5. Time Discrimination. This is measured by means of 25 units, each consisting of three notes. The first and third notes are uniformly of equal duration; namely, .74 of a second. The second note is sometimes also of the same duration, sometimes varied. The longest variation is .30 of a second and the shortest 0.3 of a second. The pupil indicates whether the three notes are of equal or different length.

6. Rhythm Discrimination. Here again are 25 pairs of rhythms. After the second rhythm, which may or may not vary in time or intensity from the first, the pupil must indicate whether the two patterns are the same or different. Difficulty is increased by increasing the number of notes in the patterns from four to eight.

7. Pitch Discrimination. This test consists of 40 tones, each played for about three seconds. In some instances the pitch is constant throughout the three seconds; in others it is varied either upward or downward. The pupil indicates whether or not the pitch remained the same.

8. Melodic Taste. In this test we find ten items, each consisting of two short melodies of two phrases. The initial phrase in each item is identical but the second always differs. Pupils are asked to give their

preference. The ten items are repeated so that pupils may make different choices if they wish but they are not allowed to change their first judgments.

9. Pitch Imagery. In administering this test a sheet containing 25 tonal patterns is placed in the hands of the students. As 25 patterns are played on the record, the pupil must indicate whether the corresponding printed pattern conforms to the one played.

10. Rhythm Imagery. The administration of this test is similar to the previous one, since pupils are given a sheet on which 25 rhythm patterns are printed. By comparing those printed with those played, they indicate whether those played are identical with those printed.

As was pointed out earlier, these tests are not devoted exclusively to the measurement of talent since the wholly untrained individual would be at a disadvantage in tests 9 and 10. It is likely that because of the more complex patterns used in many of the tests, they may be testing many other items than those listed. Yet that fact may not detract from their usefulness. While they have not had such extensive use as the Seashore tests, they may prove to be even more valuable. Percentile norms are supplied for each test for grades four to twelve and the Manual of Directions gives detailed information for administering and scoring.

A series of three tests by Max Schoen which are known as the Relative Pitch Test, Tonal Sequence Test, and Rhythm Test³ may also be classed as tests of musical talent. They have the advantage of requiring no records since the notes are produced by the examiner on a piano. Such procedure might introduce undesirable variations in the tones produced, but the tests appear to be valid.

³ Produced by The Carnegie Institute of Technology, Pittsburgh, Pennsylvania.

They are not as accessible as most tests and hence have not been widely used.

Achievement Tests. One of the best tests of musical achievement is the Kwalwasser-Ruch Test of Musical Accomplishment.⁴ Designed for grades 4 to 12, the test consists of ten items designed to measure the most common objectives of public school music education as recommended by the Music Supervisors' National Conference.

The ten subdivisions are indicated below, together with the maximum score possible on each subdivision.

1. Knowledge of Musical Symbols and Terms . . .	25
2. Recognition of Syllable Names	25
3. Detection of Pitch Errors in a Familiar Melody	25
4. Detection of Time Errors in a Familiar Melody	15
5. Recognition of Pitch Names	20
6. Knowledge of Time Signatures	20
7. Knowledge of Key Signatures	30
8. Knowledge of Note Values	15
9. Knowledge of Rest Values	15
10. Recognition of Familiar Melodies from No- tation	50
Total	240

The chief limitation of all tests of this character is that they test what the pupil knows about music—not what he can do in producing music. One may also question whether the knowledge of syllable names contributes anything since it seems likely that these names are very little used in singing songs or in any kind of musical performance except in the case of a few individuals. However, if the objectives set forth by the Music Supervisors' National Conference are valid, the test is no doubt valid also since it follows these recommendations very closely. The test

⁴ Printed by The Bureau of Educational Research and Service, State University of Iowa.

has high reliability and rather adequate norms are provided for all grades above the third. The total testing time is forty minutes.

The Gildersleeve Musical Achievement Tests ⁵ should also prove helpful in measuring the outcomes of music instruction. The tests consist of four subdivisions which require the pupils to:

1. indicate how 24 different instruments are played;
2. indicate their knowledge of musical symbols and how these are used. They are asked to locate and place "do" in eight major keys; to write the syllable names for eight tones; and to place bar lines for several measures;
3. identify famous names and a number of musical terms;
4. name ten familiar melodies from notation.

This test also appears to have good validity when comparison is made with the objectives set forth by the Music Supervisors' National Conference. Reliability is high and norms are provided for grades 4 to 8 inclusive. Testing time is about thirty minutes.

Other tests of musical accomplishment which appear to be less carefully constructed or less comprehensive are the Beach Music Test,⁶ and the Hutchinson Music Test.⁷ One might also call attention to the Hillbrand Sight-Singing Test ⁸ which is an individual test for grades 4, 5, and 6. This test consists of six short songs which the pupil is asked to sing. The examiner notes the following kinds of errors: notes wrongly pitched, transpositions, flattening,

⁵ Published by the Institute of Educational Research, Teachers College, Columbia University, New York City.

⁶ Published by Bureau of Educational Measurement, State Teachers College, Emporia, Kansas.

⁷ Published by Public School Publishing Company.

⁸ Published by the World Book Company.

sharpening, omission of notes, errors in time, extra notes, repetitions, and hesitations. The test appears to be a good one for the limited purpose for which it is designed, although it is not thoroughly objective since some errors are difficult to classify and for some examiners to detect.

Tests of Musical Appreciation. Tests of this type are extremely rare, the only one which is well known being the Courtis Standard Research Test in Music,⁹ Series M. Designed for grades 4 to 12, the test consists of two parts, one on the Recognition of Characteristic Rhythms and one on the Recognition of Mood from Melody. Victor records are used for the production of the rhythms and melodies and the pupils are supplied with printed questions in multiple choice form. Directions are read by the teacher, then the selection is played and the pupils indicate their answers. The same procedure is followed for each of the five "rhythms" and five "melodies."

To show their character, the first item of each part is reproduced below.

Recognition of Characteristic Rhythms

JOHN'S HOLIDAY

1. It was the first day of the vacation. John had decided to go to a nearby city for a holiday. The music will tell you how John made the journey. Underline the words which tell how the music says he traveled.

- | | |
|------------|------------------|
| 1. On foot | 3. On skates |
| 2. By boat | 4. On horseback. |

Recognition of Mood from Melody

JOHN GOES BLACKBERRYING

1. It was Saturday morning and the sun was shining. John's mother gave him a pail and sent him into the woods to pick berries. The music will tell you how

⁹ Published by S. A. Courtis, 1807 East Grand Boulevard, Detroit, Michigan.

John felt about going. Listen to the selection and underline the words which best express how the music says John felt.

1. He was *sorry* that he had to go.
2. He was *glad* of the chance to go.
3. He was *angry* at being sent.
4. He was *too busy* to go.

The Use of Music Tests. In view of the imperfections to be found in music tests, the user should probably be cautioned against placing too much reliance on the results obtained. As a result of their imperfections some musicians have refused to have anything to do with the tests. This attitude is as bad as that of those who rely too much on the results. In motivating certain aspects of the work music tests can be as helpful as tests in any other field. When it comes to measuring all of the outcomes, on the other hand, it seems to the writer that only a beginning has been made.

Art

The situation with respect to tests in art is not vastly different from that which obtains in the field of music. To do satisfactory art work requires a complex of abilities which do not lend themselves to measurement by simple techniques and for the most part instructors in art have not been particularly interested in the measurement of the product.

Drawing Scales. Thorndike's Scale for General Merit of Children's Drawings¹⁰ is the earliest attempt at measurement in this field and in its present form consists of 70 drawings. Of these 70 drawings, 25 are of a man; 23 of a house; 18 of a snowball fight; 1 of a boy; 1 of a girl; 1 of a

¹⁰ Published by Bureau of Publications, Teachers College, Columbia University.

horse; 1 four views of the human hands. Pupils of grades 3 to 8 are asked to draw a man or dog or house or ship or football game and the drawings are then compared with those of the scale and rated accordingly. While a certain amount of subjectivity enters into the rating thus done, the ratings do seem to be rather accurate.¹¹ No norms are provided, but the author indicates the ratings which are likely to be given drawings made by pupils in the grades. Another scale which can be used in all of the school grades is the Kline-Carey Measuring Scale for Freehand Drawing.¹² The scale is made up of four parts, one consisting of 20 drawings of a house; one of 18 drawings of a rabbit; one of 16 figures of the human form in action; and one of 19 drawings of a tree.

Scales of this type serve to call attention to desirable attributes in drawings. Especially is this true of the Kline-Carey Scale where the authors have called attention to various matters under some of the drawings. For example, under the house rated "20" we find the following:

This drawing shows the corner view of a house. It stands upright but the gable ends should slant in the same direction. The door and chimney are well placed but too small, and the chimney is crooked. When you draw houses be sure to draw the upright lines first. Draw a house that stands straight and that has straight doors, windows, and chimneys. Put it up by the side of this drawing and see if it is as good or better, and why.

*Lewerenz Tests in Fundamental Abilities of Visual Art.*¹³ These consist of nine different tests arranged in three

¹¹ Brooks, F. D., The Relative Accuracy of the Ratings Assigned with and without the Use of Drawing Scales, *School and Society* 27:518-520, April 28, 1928.

¹² Published by the Johns Hopkins Press.

¹³ Published by Southern California School Book Depository, Los Angeles, California.

booklets and are usable in grades 3 to 12. Following are the names of the tests:

- Test 1. Recognition of Proportion.
- Test 2. Originality of Line Drawing.
- Test 3. Observation of Light and Shade.
- Test 4. Knowledge of Subject Matter Vocabulary.
- Test 5. Visual Memory of Proportion.
- Test 6. Analysis of Problems in Cylindrical Perspective.
- Test 7. Analysis of Problems in Parallel Perspective.
- Test 8. Analysis of Problems in Angular Perspective.
- Test 9. Recognition of Color.

Despite the fact that scoring is in some of the tests quite subjective, the Lewerenz Tests seem to be very satisfactory for locating the abilities of elementary pupils. Norms are provided and the reliability determined by the re-test method is $r = .87$. Concerning validity the following statement is found in the author's manual.

No satisfactory criterion for validity has as yet been found. Total scores were correlated against semester grades in art with $r = .40$ (P.E. $\pm .027$). No exact knowledge is available as to how art teachers' grades correlate against themselves, but such a correlation would not be high, perhaps about .50. A correction for attenuation would obviously yield a high correlation. Judging also from experience it seems possible with the tests to rate students at the beginning of the term with nearly as great accuracy as will a teacher with a term's acquaintance behind her.

As indicated above, there are many art teachers and other persons interested in art who object to traditional methods of evaluating artistic production. The following quotation from an article by Horns¹⁴ serves to illustrate a fairly common point of view.

¹⁴ Horns, John W., *Artists Are Not Made in the Schools*, *Iowa State Teachers College Alumnus*, Volume 22, No. 4, October 1938.

Another major reason for our neglect of the art program is that the pupil, the teacher, and the superintendent are under considerable pressure to spend their best efforts on the things upon which they are to be tested and judged. There are many accomplishments of students that can be measured objectively, but art is not one of them. A teacher who knows that her work is to be judged on the basis of her pupils' marks in standard skills and information will naturally slight even the pitiful minimum of time and effort which is supposed to be devoted to art. The wholesale manufacturing of tests and measures may be an important branch of educational progress, but it does tend to over-emphasize those things which can be tested and to distort the vision of many earnest schoolmen. Why, for example, should a child be considered of low intelligence if he fails to qualify in regard to reading and information, but on the other hand shows himself to be exceptional in the matter of creative expression—possibly the field in which his energies have been directed?

To judge a student on anything less than the ultimate real value of his product is to destroy the artist. We can test objectively a child's knowledge of color schemes and even his skill in perspective. No test, however, can be devised to determine what his caricature of the teacher is worth!

Genuine art expression is always new and always unpredictable, so that no previously prepared standard is adequate to measure it.

The student, as well as the more mature artist, needs the evaluation of the teacher or critic, but that evaluation must be of a basically different sort than is commonly given. The teacher must approach each student as an individual artist and each undertaking as unique. A teacher cannot hope to give helpful criticism until he gives evidence of understanding what is being attempted. In place of more tests we need teachers who know authentic expression when they see it

and who are not bound to any formal pattern of judgment.

PROBLEMS

- 1. Assuming that measurement of artistic ability is desirable, is there any reason to believe that measurement of this ability cannot be made objectively?**
- 2. What kind of test in music appeals to you as being most helpful to the classroom teacher?**
- 3. In what ways are the problems of measurement in music and art similar? In what ways dissimilar?**
- 4. Do you agree with the point of view expressed by Horns concerning the grading of pupils on their work in art?**

BIBLIOGRAPHY

Music

- Bevins, Alice E., What Materials Shall Be Used to Teach Music in Elementary Schools? *Education* 56:536-40, May 1936.
- Fullerton, C. A., An Experiment in School Music, *Music Educators Journal* 22:26-27, March 1936.
- Jersild, A. T., and Bienstock, S. F., A Study of the Development of Children's Ability to Sing, *Journal of Educational Psychology* 25:481-503, October 1934.
- Johnson, Ruth, Making Music Live in the Elementary School, *School Activities* 7:8-9, April 1936.
- Mursell, James L., A Balanced Curriculum in Music Education, *Education* 56:521-26, May 1936.
- Mursell, J. L. and G. M., *The Psychology of School Music Teaching*, Silver Burdett Company, 1931.
- Samuelson, Agnes, The Place of Music, *School and Society* 43:825-32, June 20, 1936.
- Stanton, H., *Prognosis of Musical Achievement*, University of Rochester, New York, 1929.
- Whalen, Mary A., Music in an Activity Program, *Education* 56:531-35, May 1936.

Art

- Brooks, Fowler D., The Relative Accuracy of Ratings Assigned with and without the Use of Drawing Scales, *School and Society* 27:518-20, April 28, 1928.
- Eng, Helga, *Psychology of Children's Drawings from the First Stroke to the Coloured Drawing*, Harcourt, Brace and Company, 1931.
- Goodenough, Florence L., *Measurement of Intelligence by Drawings*, World Book Company, 1926.
- Kinter, M., and Achilles, P. S., *The Measurement of Artistic Abilities*, The Psychological Corporation, 1933.
- Lewerenz, Alfred S., Predicting Ability in Art, *Journal of Educational Psychology* 20:702-4, December 1929.

McAdory, Margaret, The Construction and Validation of an Art Test, *Contribution to Education* No. 383, Teachers College, Columbia University, 1929.

Whitford, W. G., *An Introduction to Art Education*, D. Appleton and Company, 1931.

MISCELLANEOUS TESTS FOR ELEMENTARY
SCHOOLS

Chapter IX

MISCELLANEOUS TESTS FOR ELEMENTARY SCHOOLS

Health Education

AMONG THE NEWER but most important fields of teaching in the elementary school is the field of health. That this field was for so long a time ignored was no doubt largely the result of the fact that other agencies were thought to be in charge of pupil health. With the discovery that a very large percentage of school children suffered from physical defects, it became evident that the schools must assume the responsibility, at least in part. Estimates made of the tremendous economic waste resulting directly or indirectly from physical ailments has also increased attention to these problems. While the exact objectives of health education have not been agreed upon, the following summary, by Bobbitt,¹ is at least suggestive of the many problems.

Ability to control one's dietary in such ways as to make one's food contribute in maximum measure to one's physical well-being.

Ability to keep the body mechanism properly oxygenated.

¹ Bobbitt, Franklin, *How to Make a Curriculum*, Houghton Mifflin Company, 1924.

- Ability to utilize muscular exercise as a lifelong means of maintaining a high level of physical vitality.
- Ability and disposition throughout life to engage with pleasure and profit in a varied repertory of games, sports, athletics, such as swimming, skating, hiking, rowing, riding, tennis, golf, ball games of various kinds, running games, dancing, fishing, hunting, canoeing, motoring, camping, athletic events, etc.
- Ability to carry one's self and to move and act with ease, grace, and precision.
- Ability to maintain postures conducive to the best physical functioning.
- Ability to make one's sleep contribute in maximum measure to the development and maintenance of a high level of physical vitality.
- Ability to relax physically and mentally at proper times and in proper ways.
- Ability to protect one's self from micro-organisms; and to deal with them and their products effectively in case of attack.
- Ability to take proper precautions against the spread of disease.
- Ability to dress in ways that promote the physical well-being to a maximum degree.
- Ability and disposition to maintain personal cleanliness.
- Ability to secure that variety or diversity of physical experiences necessary for maximum well-being.
- Ability to draw up an individual program of work, play, rest, sleep, meals, etc., best suited to one's physical nature and capacity.
- Ability to avoid preventable accidents.
- Ability to deal with conditions produced by many kinds of common accidents.
- Ability to care for the teeth, eyes, nose, ears, throat, skin, hair and scalp, nails, and feet.
- Ability to keep reasonably well-informed, in the degree to be expected of the layman, as to the discov-

eries of science in the fields of health conservation and promotion.

Ability to take the protective, precautionary, or remedial steps necessary to protect one's self or family from common ailments.

Ability wisely to utilize the services of physicians, nurses, dentists, and other specialists in health and physical upbuilding and maintenance.

Ability to make one's various mental and emotional states and activities contribute in maximum degree to one's physical functioning.

It is obviously difficult to measure all of these outcomes. As a matter of fact, the only tests which have been developed to any considerable extent, except for those which only physicians can employ and a few physical measurements, are tests of health knowledge. One of the best known and most comprehensive of these tests is the Gates-Strang Health Knowledge Test.² In constructing these tests a careful analysis was made of twenty selected courses in health offered in rural and city schools and of fourteen of the most widely used textbooks in the field. A total of 754 items were found that were common to at least three of these sources and were put into the form of a new-type test. The content and construction of the test were criticized by a group of experts in the field after which the test was given to several hundred school children. After an analysis of the results, some of the items were eliminated and the remaining 520 items were broken up and organized into standard tests. The topics with which the 520 items are concerned are listed below.

² Published by Bureau of Publications, Teachers College, Columbia University.

<i>Topic</i>	<i>Number of Exercises or Items</i>
Food	98
Disease Prevention	78
Physiology	33
Exercise and Posture	30
Cleanliness	27
Fresh Air and Sunlight	21
Mental Hygiene	18
Care of the Eyes	17
Safety	15
Defects, including Malnutrition	14
Clothing	14
Care of the Teeth	14
Water	13
Industrial Hygiene	12
Values of Health	12
First Aid	11
Elimination of Body Wastes	10
Care of the Hair	10
Rest and Sleep	10
Disposal of Garbage and Waste	10
Child Care	9
Names of Scientists and Miscellaneous Items	8
Alcohol	7
Care of Ears, Nose, and Throat	6
Care of the Feet	6
Patent Medicines and Drugs	5
Tobacco	4
Health Laws	3
Public Health Administration	3
Health Organization	2

Each test consists of 64 items. The directions and samples from Form I will serve to indicate their character:

Directions: Here are some questions about health. Five answers are given to each question. Read care-

fully each question and the five answers. Then mark only the *one* best answer. If you do not know which is best, mark one anyway. If more than one is marked your answer will be called wrong.

Sample Exercises

1. We should have fresh air

- .X. all of the time
- in the daytime but not at night
- at night but not during the daytime
- especially in summer
- when we begin to get a headache

-
1. It is best for children to eat

- at regular times each day
- whenever they want to eat
- six or seven times a day
- just before going to bed
- no breakfast

2. At meal time we should

- eat quickly and hurry back to school
- take a sandwich and run out to play
- eat in a hurry so we can spend more time outdoors
- sit down at the table and eat slowly
- play with the knife and fork

3. Cooked meat or vegetables should be kept

- in cool, damp cellars
- in the sunlight
- in the air, uncovered
- in a cold place, covered
- in a warm, dry closet

4. Alice is thin and pale. Besides rest and some outdoor exercise she should probably have

- candy between meals
- medicine
- meat three times a day

- more milk, bread, butter, spinach
- more fried potatoes with gravy
- 5. Flies should be kept away from food because they carry
 - soot
 - odors
 - germs
 - pollen
 - dust
- 6. There is less danger of your giving a cold to someone else if you
 - cover your face when coughing and sneezing
 - use same glass to drink from
 - lend him your handkerchief
 - exchange pencils with him
 - sit close to him
- 7. A person who has trouble with his eyes should
 - buy glasses at 10-cent store
 - borrow glasses from someone in the family
 - ask a neighbor what kind of glasses to get
 - have glasses fitted to his eyes by an eye doctor
 - not do anything about it
- 8. Of these foods the best to choose for breakfast is
 - buns
 - hot biscuit
 - pancakes
 - pie
 - hot cereal
- 9. Boys and girls who wish to be healthy should choose for a lunch at the school lunchroom
 - caramel candy, hot biscuit
 - cherry pie, doughnuts
 - nut cake and maple ice cream
 - roast pork and apple sauce
 - hot soup, baked potato, milk

10. One reason why people should not smoke is that it
- makes the skin yellow
 - makes people stupid
 - causes chills and fever
 - often harms the lining of nose and throat
 - is hard to learn how to smoke

A test of the same type by the same authors appears in the Modern School Achievement Tests. Another test of somewhat similar character is found in the Public School Achievement Tests by Orleans and Sealy.⁸ This test consists of true-false, recall-completion, multiple-choice and matching exercises, but is less comprehensive than the Gates-Strang Test. Under the title "Physiology and Hygiene" there appears in the Stanford Achievement Tests a four-response test dealing with similar subject matter.

The limitations of tests of this sort are at once apparent when one realizes that there is no perfect correlation between knowledge and activity. As we pointed out in the case of English tests, the knowledge of what is proper is only one phase of the subject. There must be instilled in the pupil a proper attitude towards matters pertaining to health, towards habits of health care; and a feeling of responsibility for his own health and the health of his family and friends. And these are most difficult to measure. Yet knowledge is basic to all other outcomes and probably much of this knowledge should be acquired in the elementary school. Tests of the knowledge thus have their place.

Elementary Science

Another comparatively new field of instruction in the elementary school and one in which objectives are not

* Published by the Public School Publishing Company.

well defined is that of elementary science. Particularly in gradation of materials is there a lack of uniformity. Some schools prefer to base their course of study on the current interests of the pupils of a given grade. It thus happens that what is discussed in the third grade in one school may be a seventh grade project in a second and omitted entirely in a third. For these reasons standardized measurement in science in the elementary school has neither progressed very far nor proved very satisfactory. Most of the known tests have been those found in the achievement test batteries such as the Modern School Achievement Tests, and the Unit Scales of Attainment, and others. As an illustration of these a portion of the science section of the Modern School Achievement Test is reproduced below.

1. An example of an animal that builds a home for its young is the
(1) beaver, (2) deer, (3) horse, (4) cow ———
2. The instrument used to find out how hot or how cold anything is, is called a
(1) barometer, (2) hydrometer, (3) hygrometer, (4) thermometer ———
3. Three of the following colors appear in the rainbow. Which one does not?
(1) red, (2) brown, (3) orange, (4) violet .. ———
4. Most plants take in water through their
(1) stems, (2) leaves, (3) roots, (4) flowers .. ———
5. Animals breathe air in order to secure
(1) nitrogen, (2) hydrogen, (3) carbon dioxide, (4) oxygen ———
6. A method by which water escapes into the air is called
(1) evaporation, (2) freezing, (3) condensation, (4) thawing ———

7. An animal that is protected by living underground is
(1) a porcupine, (2) a beaver, (3) an earthworm, (4) a red squirrel -----
8. A rock broken up into tiny bits by wind and water is called
(1) liquid, (2) air, (3) ash, (4) soil -----
9. A tree that sheds all its leaves in autumn is the
(1) pine, (2) maple, (3) cedar, (4) spruce .. -----
10. Which one of these is a flowering plant?
(1) fern, (2) moss, (3) geranium, (4) toadstool -----

It will be noted that the subjects dealt with vary greatly, not being confined to any one science or group of sciences. There is a real need for better standardized measurement in science in the elementary grades but such improvement must, no doubt, await greater standardization of courses of study and the objectives of instruction.

As in most other subjects the measurement of attitudes to be acquired as the result of science study is very inadequate. Curtis⁴ outlines the following scientific attitudes to be developed.

THE SCIENTIFIC ATTITUDES

- I. Conviction of universal basic cause and effect relations, rendering untenable
 - a. Superstitious beliefs in general, as "signs" of "good luck" or "bad luck," and charms;
 - b. "Unexplainable mysteries";
 - c. "Beats all" attitude, commonly revealed by
 1. Too ready credulity;
 2. Tendency to magnify the importance of coincidence.

⁴ *Thirty-First Yearbook of the National Society for the Study of Education, Part I, p. 56.*

- II. Sensitive curiosity concerning reasons for happenings, coupled with ideals
 - a. Of careful and accurate observation or of equally careful and accurate use of pertinent data previously collected by others;
 - b. Of patient collecting of data;
 - c. Of persistence in the search for adequate explanation.
- III. Habit of delayed response, holding views tentatively for suitable reflection (varying with the matter in hand)
 - a. To permit adequate consideration of possible options;
 - b. To permit a conscious plan of attack, clearly looking forward to a prediction of the probable outcome or solution.
- IV. Habit of weighing evidence with respect to its
 - a. Pertinence;
 - b. Soundness;
 - c. Adequacy.
- V. Respect for another's point of view, an open-mindedness and willingness to be convinced by evidence.

It is evident that while we have had some tests of the first objective, some of the others have not been tested at all.

Personality and Character Tests

Tests of personality and character have baffled several workers who have attempted to construct valid and reliable measuring instruments. There are many obvious difficulties in the way of measurement in this field, not least of which is the inadequacy of analysis of the component elements of either personality or character.

As a means of communicating to others one's ethical ideals or attitudes it is necessary to have a suitable vo-

cabulary. With this in mind Miss Schwesinger constructed a test of social-ethical vocabulary, a portion of which is reproduced below.⁵

1. BRAVERY. 1—folly, 2—courage, 3—livery,
4—impertinence, 5—humanity ——— 1.
2. SCOFF. 1—cold, 2—angry, 3—make fun of,
4—extol, 5—expound ——— 2.
3. MALICE. 1—spite, 2—poison, 3—glass, 4—
character, 5—hammer ——— 3.
4. SLUGGARD. 1—snail, 2—lazy person, 3—lax,
4—shot, 5—regard ——— 4.
5. REPROACH. 1—come near, 2—insect, 3—
scold, 4—steal game, 5—nerve ——— 5.
6. JUDICIOUS. 1—punch, 2—spoken, 3—jury,
4—wise, 5—learned ——— 6.
7. SUMPTUOUS. 1—conceited, 2—expensive,
3—repat, 4—meager, 5—fairlylike ——— 7.
8. INTROSPECTIVE. 1—look over, 2—inspec-
tion, 3—self-examining, 4—inward, 5—in-
sight ——— 8.

In addition to a suitable vocabulary, other knowledge is necessary. For testing this Hartshorne and May have developed a Cause-Effect Test⁶ which they illustrate as follows.

Some of the statements made below are true and some are false. Read each statement carefully and underline the word TRUE if it seems to you to be true. Underline the word FALSE if it seems to you to be false.

Good marks are chiefly a matter of luck	True	False
Ministers' sons and deacons' daughters usually go wrong	True	False
If one eats stolen apples, he will have a stomach ache	True	False

⁵ Hartshorne and May, *Studies in the Organization of Character*,
p. 43.

⁶ *Ibid.*, p. 38.

Success always comes from hard work .	True	False
God punishes bad people by making them sick	True	False
Eavesdroppers never hear anything good about themselves	True	False
The youngster who can cheat and not get caught at it shows more good sense than the one who does not cheat	True	False

Another type of knowledge test is the Recognition Test⁷ illustrated below.

After each statement are five letters: C, L, S, X, J. If the deed is a case of cheating, draw a circle around the C; if it is lying, around the L; if it is stealing, around the S. If it is something wrong, but not either cheating, lying, or stealing, put a circle around the X. If it is not wrong at all, put a circle around the J. If the thing is both cheating and lying or stealing and lying or all three, encircle all the letters you need to in order to express your opinion.

Bullying younger children	C.	L.	S.	X.	J.
Using street car transfers that are out of date	C.	L.	S.	X.	J.
Riding on the back of a truck without the driver's knowing it	C.	L.	S.	X.	J.
Apologizing for a misdeed when you are not really sorry	C.	L.	S.	X.	J.
Forgetting to brush your teeth for a day	C.	L.	S.	X.	J.
Talking loudly in the hallways when classes are in session ..	C.	L.	S.	X.	J.
Picking flowers in a public park	C.	L.	S.	X.	J.
When you don't want to go somewhere, making up an excuse so as not to hurt anyone's feelings	C.	L.	S.	X.	J.

⁷ Hartshorne and May, *Studies in the Organization of Character*, p. 40.

Still another test of knowledge is known as the Probability Test ⁸ offering alternatives as indicated.

This is likely to happen	This might happen, but not likely	This would not happen	John started across the street without looking both ways
---	---	---	1. He got hit with an automobile.
---	---	---	2. He caused an accident to other people.
---	---	---	3. The traffic laws were changed so that boys could cross more safely.
---	---	---	4. It scared the automobile drivers who saw him.
---	---	---	5. He was the cause of an automobile driver being put in jail.
---	---	---	6. Two cars collided trying to avoid him.
---	---	---	7. He got across as safely as anyone.
---	---	---	8. He got confused when he looked up and saw a car coming.

The Duties Test,⁹ in which the subject indicates whether or not it is his duty to do the things mentioned, is another of these tests. In indicating his response, the pupil underlines "Yes" or "No" or "S" (for sometimes "Yes" and sometimes "No"). Some of the items are:

⁸ Ibid., p. 45.

⁹ Ibid., p. 46.

- | | | | |
|--|-----|---|----|
| 1. To help a slow or dull child with his lessons | Yes | S | No |
| 2. To read the newspapers every day .. | Yes | S | No |
| 3. To call your teacher's attention to the fact if you received a higher grade than you deserved | Yes | S | No |
| 4. To keep a diary | Yes | S | No |
| 5. To sneeze when you feel like it | Yes | S | No |
| 6. To jeer at a child who has just been punished | Yes | S | No |
| 7. To smile when things go wrong | Yes | S | No |
| 8. To report another pupil if you see him cheating | Yes | S | No |

The Provocations Test,¹⁰ so called because the situations named are provocative of responses that are in conflict with ideal modes of response, appears to demand more careful discrimination. It is similar in form to the Duties Test, as shown below:

Here are some little stories of what some children did. You are to decide whether they did right or wrong. If what they did was not quite right, perhaps it was at least excusable in view of the circumstances. Look at the sample first.

Sample: Jane's family were too poor to buy fruit for her sick brother. So every now and then Jane took an apple or an orange from a fruit stand and brought it home to him.

Now if you think she was absolutely wrong in taking the fruit, put a circle around the Wr, like this ...

R Ex (Wr)

But if she did exactly right, encircle the R, like this

(R) Ex Wr

If you think she was wrong but excusable in view of her desire to bring it to her sick brother, encircle the Ex, like this

R (Ex) Wr

¹⁰ Hartshorne and May, *Studies in the Organization of Character*, p. 50.

Begin here and do the rest in the same way:

1. Helen noticed that nearly everyone in the class was cheating on a test; so she cheated too R Ex Wr
2. Harry was a Christian boy. One day a Jewish boy called Harry a "dirty Christian." Harry knocked him down R Ex Wr
3. Charles did not want to play marbles for keeps, but the boys called him a "sissy"; so he went ahead and played for keeps anyway R Ex Wr
4. On the way to Sunday School Jack matched pennies with the other boys in order to get some money for the Sunday School collection R Ex Wr

Tests of attitude have been somewhat difficult to construct, but much more difficult to validate for the reason that a pupil's response to a given question does not indicate what he would actually *do* if a similar situation presented itself. More than that, his response does not even indicate what he thinks one should do; rather it indicates what he believes others (particularly the examiner) think should be done under the circumstances. Perhaps there is some positive relationship between his real attitudes and his attitudes as indicated in a pencil and paper test, but up to this time we do not know the extent of that relationship.

Among these tests of attitudes is a series by Betts known as The Northwestern University Citizenship Tests. In one portion of the test the pupil is asked to indicate how he would feel if his best friend did certain things. Another portion is devoted to indicating the seriousness of a series of actions. One section of this test is known as "The Citizen at Home" test. The items on the following page indicate its character.

EXTRACT FROM NORTHWESTERN UNIVERSITY CITIZENSHIP TESTS—THE CITIZEN AT HOME

Ballot 1. How Serious (Wrong) Is It?

Theodore Roosevelt once scornfully called some men who had done things he thought were wrong "undesirable citizens." What is it that makes one a good or a bad citizen in his home? Would you not say that it is his acts, the way he conducts himself?

Some of the items (acts) listed below are probably more important than others in making good citizens in the home. You may think some of the acts are serious, or wrong, and that others are not. Vote in the proper column for each item to show how serious or wrong the act is. Vote but once for each item. Think carefully, vote as you think. Make the best score you can.

	<i>very, serious</i>	<i>some- what serious</i>	<i>not very serious</i>	<i>not serious at all</i>
.....				
.....				
.....				
.....				
.....				

How Serious (Wrong) Is the Act? I Think It Is—

- Joyce has the habit of coming late to her meals quite often.
- William tells his mother he has forgotten to do her errand when he really has not forgotten, but has spent the time practicing baseball with his team.
- Finding ten cents on the bureau, Elsie takes it without asking anybody and buys a pencil which she needs.
- When getting ready for bed, Lloyd lets his shoes drop to the floor as he takes them off, rather than placing them quietly on the floor.
- Tom's father gave him a half-dollar for his school bank. On his way to school Tom sees in a store window a baseball marked down from one dollar to fifty cents, so he buys the ball with the half-dollar, meaning to earn money soon and then put the half-dollar in the bank.

A number of ingenious attitudes tests have been devised by Hartshorne and May. One of these known as the Punishments Test asks pupils to indicate whether they consider various offenses grave or not by designating the degree of punishment which should be meted out. A portion of this test is reproduced here.

PART I		Do Not Punish	PUNISH					
			<i>Very lightly</i>	<i>Lightly</i>	<i>Rather strictly</i>	<i>Hard</i>	<i>Very severely</i>	
1.	Looking up answers in a book during an examination							1.
2.	Peeping in a game of blind man's buff at a party							2.
3.	Tripping and dirty play in a basketball game							3.
4.	Taking some candy from the teacher's desk							4.
8.	Copying from another pupil's paper on an examination							8.
12.	Pretending not to hear when someone calls							12.
13.	Refusing to help buy a radio for the class							13.
14.	Reading a story during study period							14.
20.	Refusing to help make toys for sick children							20.
25.	Keeping an article you found when the owner's name is on it							25.

In addition to the tests mentioned above, a number of clever devices have been used to determine what a pupil will do when opportunities are present for certain forms of deception. Voelker and Cady used a technique whereby they could discover to what extent a pupil would improve his own score on a test when allowed to score his own paper. Other techniques employed by Voelker include the following.¹¹

1. *The Over-Change Test.* The subject is sent on a purchasing errand. It is prearranged with the merchant to give him a certain amount of over-change. The test is what he will do with it.

2. *The Let-me-help-you Test.* The subject is given a difficult task such as a puzzle, which he promises to do without receiving help. A confederate of the examiner incidentally offers help. If the subject refuses it, he is scored plus; if he accepts, the score is minus. A variation of this is to give the pupil a simple arithmetic test, telling him not to look on the back page. On this page is printed a series of answers, some of which are right and some wrong. If he disobeys and copies from the back sheet, he is likely to copy some of the wrong answers and thereby get caught.

3. *The Reliability Test.* The subject agrees to deliver a letter to his parents and see that a reply is mailed within twenty-four hours. He further agrees not to read the letter. The letter is left unsealed. It is a letter to the parent about the bearer, asking for ratings on certain character traits. If no reply is received or if the reply appears to have been written by the bearer, he is scored as deceptive (untrustworthy).

4. *The Missent Letter.* The subject receives a letter from a business firm enclosing twenty-five cents. The letter says that this amount is sent to balance his account with the firm, and requests that the receipt be

¹¹ Hartshorne and May, *Studies in Deceit*, p. 44.

sent back in the addressed and stamped envelope which is enclosed.

Hartshorne and May have added to these tests a number which indicate how much a person will take advantage in playing a game, working a puzzle, seeking to win approval, or avoid disapproval, and in athletic performance.

While tests of this type do give an index to one's reactions in situations similar to these, they have decided limitations. In the first place, they are confined to a single situation, and honesty in one situation does not necessarily indicate honesty in another which may appear quite similar. The individual's "mental set" at one time might well be such that, for example, being in desperate need of money, he would stoop to any form of stealing, whereas on another occasion, when this need is past, he would withstand almost any form of temptation. Similarly, in avoidance of disapproval or in cheating in tests, varying temporary attitudes cause differences in conduct. This does not mean that attempts at measurement should be scorned. Certainly the development of the common virtues and of proper attitudes is an important objective of education and until valid measures of these important factors are available, we shall not know when they are attained or how pupils most readily attain them.

The Character Education Inquiry, under the direction of Hartshorne and May, has revealed some interesting facts concerning character. Those who are particularly interested in this field will read with profit the reports of this inquiry. The first three volumes are called *Studies in Deceit*, *Studies in Service and Self-Control*, and *Studies in the Organization of Character*.

Torgerson's Pupil Adjustment Inventory. A major contribution in the field of pupil adjustment has been made

by T. L. Torgerson in his *Pupil Adjustment Inventory*.¹² The pupil who fails to make the proper adjustment to the schoolroom situation, whether it be in the matter of proper application to his studies, his attitude towards his teacher or fellow pupils, his acceptance of responsibility, emotional control, or overt conduct, is a serious menace to the establishment of proper schoolroom procedures. More serious still is the fact that these maladjustments may contribute toward serious personality difficulties or mental disturbances that carry over into his life when the period of formal education is completed. To note such maladjustments, to determine their causes and, in so far as possible, to apply remedial treatment is a function of the teacher which all too frequently has been neglected. It is to aid the teacher in the first two of these functions and to suggest effective remedial treatment that these materials have been devised.

Part A of these materials consists of a *Pupil Adjustment Inventory* on which one or more teachers rate the pupil on social attitude, emotional control, nervousness, day dreaming, responsibility, interest, laziness, happiness, conduct, and success in school. Part B consists of a case study record in which the teacher may conveniently record a great many data which may have significance for the understanding of the pupil being studied. These are listed under the major headings of pedagogical factors, including school history and study habits; physical factors; emotional and social factors including interests, attitudes, and home environment; intelligence test ratings; achievement test ratings; aptitude test ratings; and personality and adjustment ratings. Several pages of the booklet are also provided for recording the history of the case, the remedial treatment used, and the results of treatment.

¹² Published by E. M. Hale and Company, Milwaukee, Wisconsin.

Part C is devoted to an exposition of common symptoms and causes, and to a discussion of seventy-five remedial procedures which have been found effective with various types of cases. With such materials the teacher should be able to diagnose and treat practically all of the cases of maladjustment found in the ordinary schoolroom though for a few incorrigibles and confirmed problem cases it may be necessary to call in the assistance of psychologists, psychiatrists, or other technically trained educationalists.

PROBLEMS

1. List some means employed in elementary schools for creating proper attitudes towards matters pertaining to health.
2. Should the course of study in science be organized so as to be uniform throughout a given school system or should it be dependent upon the expressed interest of the pupils?
3. List the various characteristics that go to make up one's personality. How do you judge the personality of an acquaintance?
4. Can you give illustrations which indicate that a so-called "honest" person has different standards for different occasions?

BIBLIOGRAPHY

- Chenoweth, L. B., and Selkirk, T. K., *School Health Problems*, F. S. Crofts and Company, 1937.
- Fitzpatrick, F. L., Pupil Testimony Concerning Their Science Interests, *Teachers College Record* 38:381-88, February 1937.
- Forsythe, W. E., and Rugen, M. E., Health Knowledge Test, *American Physical Education Association Research Quarterly* 6:105-20, May 1935.
- Hartshorne, H., and May, Mark A., *Studies in Deceit*, The Macmillan Company, 1928.
- Hartshorne, H., May, Mark A., and Maller, J. B., *Studies in Service and Self-Control*, The Macmillan Company, 1929.
- Hartshorne, H., May, Mark A., and Shuttleworth, Frank K., *Studies in the Organization of Character*, The Macmillan Company, 1930.
- Jones, J. H., Comparison of Health Knowledge and Health Instruction at the Sixth Grade Level in Certain Rural and Urban Schools, *American Physical Education Association Research Quarterly* 6:53-60, May 1935.
- Lacy, Frances, *Building Health Through a School Activity*, *Childhood Education* 12:273-75, March 1936.
- LuPone, O. J., Some Problems That Must Be Answered in Elementary Science, *School Science and Mathematics* 38:666-72, June 1938.
- National Society for the Study of Education, A Program for Teaching Science, *Thirty-First Yearbook*, Part I, Public School Publishing Company, 1932.
- Norton, J. K., and Norton, M. A., *Foundations of Curriculum Building*, Chapter 5, Ginn and Company, 1936.
- Robertson, Martin L., The Selection of Science Principles Suitable as Goals of Instruction in the Elementary School, *Science Education* 19:1-4 and 65-70, February and April 1935.
- Rogers, F. R., Notable Achievement in Health Education, *Education* 55:317-19, January 1935.

- Silver, Harry B., Adequate School Health Programs, *Educational Administration and Supervision* 24:303-312, April 1938.
- Traxler, Arthur E., *The Use of Tests and Rating Devices in the Appraisal of Personality*, Educational Records Bureau, 1938.

GENERAL TESTS OF ACHIEVEMENT

Chapter X

GENERAL TESTS OF ACHIEVEMENT

AS HAS BEEN POINTED OUT in previous chapters, various test makers employ a variety of methods of expressing test scores. As a result it is in some instances difficult to make comparisons between the results of two tests so as to determine whether pupils have progressed during the interval between tests. Even when the same terminology is employed in expressing derived scores, there is still the not remote possibility that different tests have, by reason of standardization on groups of non-similar pupils, quite different norms. A test standardized on schools in the well-to-do districts of large cities might, for example, have very different norms for the same accomplishment from those standardized on schools in poor mining communities. Most test makers, of course, attempt to establish norms by administering their test to pupils of all levels, but not infrequently they fail to secure a truly representative sampling despite efforts to do so. Similarly, if a teacher wishes to compare the reading ability of her pupils with their ability in arithmetic, there is a danger that the norms for the reading test are not comparable with those for the arithmetic test.

It sometimes happens also that a teacher or administrator is primarily interested in the general achievement of pupils without particular reference to their achievement

in any given subject. In order to secure this information, it is not expedient to average the scores on various tests since by doing so one may be giving undue weight to one subject and slighting others. By the use of Z-scores, B-scores, or T-scores, this difficulty may be overcome.

Another method of overcoming the difficulties mentioned above is by means of employing so-called test "batteries" which test in all or most of the elementary school subjects and which have been standardized by the same method and with the same pupils as subjects. Such batteries have been rather commonly used, chiefly because of their convenience, and a number of them, including those described below, have been published in recent years.

*The Otis Classification Test*¹

This test is representative of those general tests in which little or no attempt is made to determine what the pupil has achieved in each of the subject fields separately but where interest is centered simply in a composite of achievement. In addition, this test includes a mental test as the second section, which is really the Otis Self-Administering Test of Mental Ability mentioned in a later chapter. The items in the achievement section are arranged in cycle form, the first cycle containing the easiest items, the second those of medium difficulty and the third the most difficult items.

It is obvious that this test cannot be recommended for the purpose of determining a pupil's standing in any one subject since there are only 7 items in physiology and hygiene; 4 in literature; 12 in grammar and dictation; 5 in music, etc. Such a small number of items could hardly give a reliable index to a pupil's achievement in a

¹ Published by World Book Company.

given field regardless of the care with which the items are chosen. The test serves its chief usefulness in classifying pupils for instructional purposes, either for homogeneous grouping or for grade placement. For the latter purpose other tests are probably more effective.

There are three forms of this test. Each part is given in thirty minutes, exclusive of the time for administration. The test is designed for grades 4 to 8, but the author states that it has also been used successfully in grades 3 and 9. Scoring is entirely objective and is simplified by the use of convenient scoring keys.

*The New Stanford Achievement Tests*²

This well-known battery devised by Kelley, Ruch and Terman is a revision of the earlier Stanford Achievement Tests by the same authors. The Primary Examination is designed for grades 2 and 3, and the Advanced Examination, for grades 4 to 9. They consist of series of tests in the individual school subjects and some of them have been discussed in connection with the various subject tests.

The Primary Examination is composed of two tests in reading, one in spelling (dictation), and two in arithmetic. The reading test consists of a paragraph reading section and a section devoted to vocabulary. The arithmetic test is also divided, one part consisting of reasoning problems, the other of computational problems. It is recommended that this examination be given in two sittings, of twenty-one and forty-two minutes respectively.

The Advanced Examination consists of ten tests. The names of the tests are given below and the working time on each is indicated.

² Published by World Book Company.

<i>Test</i>	<i>Working Time</i>
1. Reading: Paragraph Meaning	25 min.
2. Reading: Word Meaning	10 min.
3. Dictation (Spelling)	about 15 min.
4. Language Usage	10 min.
5. Literature	10 min.
6. History and Civics	10 min.
7. Geography	10 min.
8. Physiology and Hygiene	10 min.
9. Arithmetic Reasoning	20 min.
10. Arithmetic Computation	30 min.

Four sittings are recommended for this test to avoid undue strain on the pupils.

The scoring is thoroughly objective except for the Paragraph Meaning Test and even here there is no serious question as to how an item should be scored. Despite the use of scoring keys, considerable time is required in scoring a complete test, from ten to fifteen minutes ordinarily, even by experienced workers. A very convenient device is employed, however, for the conversion of scores into educational age and grade equivalents. At the close of each test there appears at the bottom of the page a scheme like that illustrated below which is taken from the Arithmetic Reasoning Test.

No.	Rt.	0	1	2	3	4	5	6	7	...	34	35	36	37	38	39	40
Score	3	10	17	24	31	37	42	49	...	119	120	122	124	125	127	129	

The score on any test is thus made equivalent to the score on any other test and by transferring the score to the inside cover page the pupil's educational profile may be shown.

While the authors are aware of this situation, it is doubtful whether the typical classroom teacher who uses such a test will understand that the tests in literature,

history and civics, geography, and physiology and hygiene cannot be relied upon for individual measurement. Other dangers in composite achievement tests, with particular reference to the older Stanford Test, are pointed out by Wilson and Hoke ³ as follows:

Manifestly, the value of the composite achievement test is dependent upon the values of the various parts of the test. This means, therefore, that to parts of the test dealing with reading must be applied the same principles and standards applicable to any other good reading test. To the arithmetic part of the test must be applied the same standards applicable to any good test in arithmetic. The same may be said of the test in language. There is always the difficulty in attempting to cover whole fields in a single test, that the authors may not have the special equipment required in each of the several fields. There is some evidence of this in the Stanford test. In arithmetic, for instance, there are a number of problems which go beyond reasonable social usage. Pupils are required to add decimals and common fractions on a basis which would be difficult, if not impossible, for the average adult. Pupils are asked to perform operations in subtraction, addition, and multiplication of compound numbers, and studies on social usage have shown that these processes have practically zero value.

The Modern School Achievement Tests

A part of the above criticism by Wilson and Hoke is not pertinent when applied to the Modern Tests,⁴ since these have been prepared by the following specialists: Arthur I. Gates, Paul R. Mort, Percival M. Symonds, Ralph B.

³ Wilson and Hoke, *How to Measure (Revised)*, p. 227, The Macmillan Company. Quoted by permission.

⁴ Published by Bureau of Publications, Teachers College, Columbia University.

Spence, Gerald S. Craig, De Forest Stull, Roy Hatch, Amy I. Shaw, and Laura B. Krieger.

The sub-tests in this battery as well as time limits and recommendations for division into "sittings" are given in the following summary from the authors' Manual.

<i>First sitting:</i>	<i>Minutes</i>	
Distribution of papers, filling out		
page 1	5	
Reading Comprehension	30	
Reading Speed	5 or 8	40 or 43
<i>Second sitting:</i>		
Arithmetic Computation	20	
Arithmetic Reasoning	25	
Directions	1	46
<i>Third sitting:</i>		
Spelling	(about) 15	
Health Knowledge	15	
Language Usage	10	
Directions	2	42
<i>Fourth sitting:</i>		
History and Civics	15	
Geography	15	
Elementary Science	12	
Directions	3	45

A number of these sub-tests have also been discussed in other connections.

Concerning the validity of the tests, the authors make the following statement.

The first problem in the selection of subject matter for the tests was to obtain materials which could meet high standards for curriculum content without violating standards of objective measurement. Tests exercise considerable influence both on the content of the curriculum and on methods of teaching, and the items must therefore represent only the best curricu-

lum materials. Three checks helped to guarantee that this requirement was met: (1) The experience of the authors in the field. Active participation of the authors in the development of curriculum material, and their wide acquaintance with best practice in the field justify confidence of test users in the worth of the materials selected. (2) Comparison with best curriculum practice. Courses of study selected as the best in the country by the careful rating of the Bureau of Curriculum Research of Teachers College, Columbia University, were studied. Only materials found in these courses of study were retained. (3) Checking against best teaching practice. The items were submitted to master teachers in different parts of the country and unacceptable items were then discarded.

The following statement concerning norms is quoted because it seems to the writer of this volume to be pertinent not only to norms for this particular test, but to norms in general.

The norms or standards for the tests were established on the basis of the scores of 6710 children in 37 cities. The criticism often advanced against standard tests because of the blind use of norms as an absolute criterion of excellence can be obviated, in part at least, by proper consideration of the numerous factors which determine acceptable achievement. It is planned in a subsequent pamphlet to publish the results obtained by a number of cities which deviate more or less widely from the established norms for all cities. The educational programs of these cities, to be selected because of their reputations for progressive education, are undoubtedly as sound and defensible as those of any other cities. Their records will be given to illustrate the direction and amount of deviation from individual test norms to be expected from schools with progressive programs of education. Variations in test scores are as frequently related to the amount of time allotted to different subjects and to

emphasis placed by supervisors on different subjects as they are to the quality of teaching. The question of allotment of time, or the proper weighting to be given to any of the other factors which influence test results, must be answered by each community in the light of its own conditions and educational aims. The mere fact of variation from a country-wide average is no necessary indication that an individual city does not have a well-balanced program, or that it is not carrying on effective teaching. School authorities in cities so varying are of course interested in the amount and direction of such variation as a check on their objectives. It is believed that the publishing of multiple norms will help to encourage exchange of ideas among schools, stimulate efforts to appraise test results in terms of all local factors that are significant, and help discover trends in the placing of emphasis on individual subjects.

*The Unit Scales of Attainment*⁵

The tests discussed above are designed for use in a number of grades. As we have previously pointed out, a test designed for several grades is likely to contain a considerable amount of material which is non-functioning for pupils of a given grade. For example, the amount of non-functioning material in the New Stanford Test for a fourth grader or a ninth grader must be rather large. For the fourth grader, much is too difficult to attempt; for the ninth grader, much is so easy that it merely consumes his time without really testing. In an apparent attempt to limit the amount of non-functioning materials the authors of the Unit Scales of Attainment have produced separate tests for grades 3 and 4, for grades 5 and 6, and for grades 7 and 8.

⁵ Published by Educational Test Bureau, Inc., Minneapolis, Minnesota.

The tests for grades 5 and 6, as well as those for grades 7 and 8, contain the following parts:

<i>Tests</i>	<i>Time Limits</i>
Reading	45 min.
Geography	10 min.
Literature	9 min.
Elementary Science	9 min.
American History	12 min.
Arithmetic Problems	22 min.
Arithmetic Fundamentals	20 min.
Spelling	(about) 10 min.
English—Capitalization	(about) 10 min.
Punctuation	(about) 10 min.
Usage	(about) 10 min.

Scores are converted into C-scores. In describing these scores the authors state:

Since the Scales are like yardsticks with a task of known value in place of the inch mark at each unit distance, the C-scores measure amounts of abilities just as numbers of inches measure amounts of height or numbers of pounds measure amounts of weight. The C-score unit is approximately one-tenth of the quartile of all pupils of the same chronological age, or $1/68$ of the difference between the poorest and best pupil in each subject among a sampling of 100 pupils of the same chronological age representative. The C-score is not only expressed in terms of a consistent unit of measurement throughout the range of the scales but is an unweighted score in that it has the same meaning at every point on the scale. It shows just how difficult tasks a pupil can attempt and succeed in half of the tasks. The level of difficulty at which half of the tasks can be done was selected as the one in which to express the C-score because at this point the error of measurement is smaller than at any other percentage of correctness. The number right on an unscaled test, not only gives no clue as to the proportion of tasks

that can be successfully done by the pupil, but the proportion even varies with the different scores. It is the consistency as well as the definiteness of meaning of the C-score that makes feasible the comparison of gains made at all points on the scales and also from one scale to another. Thus the gain made by a pupil in **READING** in going from a C-score of 60 to one of 66 is the same as the gain made by another pupil in going from a C-score of 80 to one of 86 in **READING** or a gain made from a C-score of 70 to one of 76 in **GEOGRAPHY**. The C-scores express the amounts of abilities with reference to a definite zero point that is 60 C-score units below the ability of normal children mentally ten years of age. A pupil who makes a C-score of 90 in the **FUNDAMENTAL OPERATIONS** of arithmetic has thus gained approximately twice as much beyond the normal C-score of 60 at ten as he has gained in **GEOGRAPHY** if he makes a C-score of 75.

Reliability coefficients are not reported, but the following quotation from the authors' Manual is pertinent.

For any single scale included in the Unit Scales of Attainment the probable error of measurement lies between 2.5 and 3.5 C-score points. This means that the chances are even that the pupil's C-score in any scale does not vary from his true ability by more than 2.5 to 3.5 C-score points, the C-score unit approximating very closely one-tenth of a quartile for pupils of the same chronological age. The probable error of measurement for the individual scales is thus about one twentieth of the difference between the best and poorest pupil among a representative sampling of one hundred pupils of the same chronological age. This approximates the gain normally made during half a year in school. While this may seem to be a considerable error of measurement there is no way of reducing it for an educational measuring instrument containing no more items or taking no more time than these scales except by narrowing the function measured,

which would not be desirable; introducing a time limit for all pupils which would render the scores as meaningless as those on an ordinary test, or matching the items for content, which would render the scores merely measures of the knowledge of the content of the particular test used.

*The Metropolitan Achievement Tests*⁶

The authors of this series of tests also recognize the desirability of restricting the number of grades in which a test is to be used. There are four different batteries, as follows:

Primary I Battery, for Grade One, consisting of tests in word and phrase recognition, word meaning, and numbers. Total time, 60 minutes.

Primary II Battery, for Grades Two and Three, consisting of tests in reading completion, paragraph reading, vocabulary, arithmetic fundamentals and problems, language, and spelling. Total time, 90 minutes.

Intermediate Battery, for Grades Four, Five, and Six, consisting of tests in spelling, reading, vocabulary, arithmetic fundamentals and problems, English, literature, history, and geography. Total time, 3 hours and 40 minutes.

Advanced Battery, for Grades Seven and Eight, consisting of tests in the same fields as those tested by the Intermediate Battery. The total time is 4 hours.

Considerable information is furnished in the Supervisor's Manual concerning the methods used to insure the validity of the items. In summarizing their statement concerning validity, the authors say: "The content, in the first place, is based on courses of study, research studies of content and widely used textbooks. The selection of questions has been checked carefully to insure satisfactory

⁶ Published by World Book Company.

representation of grade content—among the several grades and within each grade. All necessary statistical procedures were employed to insure adequate statistical bases for the selection of questions. The wording of questions, the vocabulary employed, and the directions for administering and scoring were carefully prepared and checked in the light of pupil responses and criticisms made by teachers and supervisors. In some instances items were repeated (for example, types of problems in arithmetic) where necessary to do so to make the test more valid.”

Scoring of most of the tests is completely objective and fairly rapid, scoring keys in convenient form being provided with the tests. Scores are readily converted into grade equivalents by means of a device not unlike that used in the Stanford Achievement Tests.

*The Progressive Achievement Tests*¹

An attempt to include in a general achievement test materials which may be used not only for survey purposes but also for more or less detailed diagnosis has been made by Tiegs and Clark. The Primary Battery is for grades 1, 2, and 3; the Elementary Battery for grades 4, 5, and 6; the Intermediate Battery for grades 7, 8, and 9. At each level one finds tests in Reading Vocabulary, Reading Comprehension, Arithmetic Reasoning, Arithmetic Fundamentals, and Language. The various subdivisions differ with the different batteries. Thus the sub-tests under Reading Vocabulary are, in the Primary Battery: Word Form, Word Recognition, and Meaning of Opposites. In the Elementary Battery they are, for the same test: Word Form, Word Recognition, Meaning of Opposites, and

¹ Published by Southern California School Book Depository, Los Angeles, California.

Meaning of Similarities; while in the Intermediate Battery they are: Mathematics, Science, Social Science, and Literature.

Reliability coefficients reported are based on data from two or three grades but reports for single grades cover only grades 2 and 3. All of the coefficients reported are above .86 for each subject test, when a two or three grade range is used. While these coefficients are high, they do not appear to be higher than those of the corresponding sections of other good batteries. Perhaps one reason that they are not higher is the fact that, in keeping with the usual practice for diagnostic tests, fixed time limits are not used for most tests. The directions simply indicate that when about ninety per cent of the pupils have finished, the examiner is to have the pupils stop or turn to the next test.

Concerning the validity of the Elementary Battery, the authors make the following statement: "In determining the content of the test battery, test situations were selected which represent the essential elements of the basic skills which are being taught in grades four, five, and six. The materials were selected to measure the types of abilities which are indicated as desirable educational objectives in recent courses of study and are in accordance with progressive educational practice." Similar statements are made concerning the other batteries.

Under the heading of "Language" there is an attempt to measure handwriting. A scale is used for scoring the words written for spelling. Such a practice has one advantage; namely, that the pupils are more likely to write in their customary style when they are not aware that writing is being tested.

Except for the handwriting, the scoring is objective and keys are well arranged for rapid checking.

If we follow the terminology of this volume, we should probably say that the tests are analytical rather than diagnostic, since some of them are too limited for any complete diagnosis. For example, the Words and Sentences section of the Language Test (Intermediate Battery) contains but ten items for word usage and a like number of statements which are complete or incomplete sentences. Similarly, the Spelling Test in the same battery calls for the spelling of only thirty words.

Perhaps the greatest difficulty in the use of this battery is the fact that by no means all of the subject fields are tested. Except for the vocabulary in some of these fields, there is no test in history, geography, literature, health, or elementary science. Perhaps the authors felt that they would rather deal only with those subjects in which relatively good diagnostic tests could be constructed, omitting those where, because of great differences in courses of study, good tests are difficult to construct.

Self-Marking Achievement Batteries

Another series of achievement tests which are about to come from the press is in the Self-Marking series⁸ of tests. These consist of four batteries, one for the first grade including a test in reading and one in number concepts; one for grades two and three including tests in reading, number, English and spelling; one for grades four, five, and six; and one for grades seven, eight, and nine. The batteries for grades four and higher include sub-tests in reading, English, arithmetic, spelling, social studies, science, and geography. Each test is constructed by one or more authorities in the field and some of them contain unique features.

⁸ Published by Houghton Mifflin Company

While norms are well established by administration in a considerable number of widely scattered schools, the use of the tests has as yet been restricted because up to the present time they have not been placed on the market. One of the most attractive features is, of course, the rapid scoring which can be done because of the use of the self-marking device. Studies have revealed that the scoring of one battery not in this series could be done at the rate of five or six per hour, while the scoring of these batteries can be done at the rate of from thirty to sixty per hour.

Sangren Information Tests for Young Children ⁹

This is one of the most satisfactory tests of pre-school achievement but is less widely used than some because it must be given as an individual test, thus consuming much time. Designed for kindergarten and first grade, the test is divided into six parts, each part dealing with information which children often acquire before coming to school. The six parts are: Nature Study, Numbers, Vocabulary, Social and Civic Information, Household, and Language and Literature.

The chief use of this test appears to be for grouping children for instructional aid. For this purpose it appears to be at least as useful as the Stanford Revision of the Binet-Simon Test, and agrees with this test in most instances. The author points out that the correlation between this test and chronological age is very low, whereas this correlation in the Stanford-Binet Test is much higher. Since mental age is admittedly a better index of school success than is chronological age, it appears that this test also is a better measure than is chronological age for grouping children or for placing them in proper grades.

⁹ Published by World Book Company.

Coefficients of reliability are reported as .975 when kindergarten children are the subjects and .961 when first grade pupils are examined. All in all, the test has many commendable features which would, no doubt, cause it to be used more generally were it not for the time required for administration. The author recommends giving only three of the six tests at one sitting.

PROBLEMS

1. In what situations would one prefer to use a battery of tests such as the New Stanford Achievement Tests, Unit Scales of Attainment, etc.? For whom are they most useful, the classroom teacher or the supervisor?
2. Compare the usefulness of tests of general achievement with such tests as the Nelson Silent Reading Test or the Compass Tests in Arithmetic.
3. How do batteries of tests fit into the diagnostic and remedial program?
4. Secure samples of several test batteries and examine them carefully with a view to recommending one of them to an elementary school principal for use in grades 4, 5, and 6. List reasons for your recommendation.

BIBLIOGRAPHY

- Gilliland, A. R., Jordan, R. H., and Freeman, Frank S., *Educational Measurements and the Classroom Teacher*, Revised, Chapter 16, The Century Company, 1931.
- Greene, H. A., and Jorgensen, A. N., *The Use and Interpretation of Elementary School Tests*, Chapter 20, Longmans, Green and Company, 1936.
- Webb, L. W., and Shotwell, Anna M., *Standard Tests in the Elementary School*, Chapter 18, Ray Long and Richard Smith, Inc., 1932.
- Wilson, G. M., and Hoke, K. J., *How to Measure*, Revised, Chapter 9, The Macmillan Company, 1928.

THE MEASUREMENT OF INTELLIGENCE

Chapter XI

THE MEASUREMENT OF INTELLIGENCE

ONE OF THE MOST INTERESTING CHAPTERS in the history of the testing movement relates to the development of adequate tests of intelligence. Before entering upon the history of the movement, however, let us first discuss briefly certain controversial points concerning the nature of intelligence and its development.

Controversies Concerning Intelligence. What is intelligence? Many definitions of intelligence have been offered and some of them differ widely from others. Some years ago the *Journal of Educational Psychology* invited several of the leading educational psychologists to submit their definitions of intelligence for publication in that periodical. We may take as Colvin's definition the following quotation: "An individual possesses intelligence in so far as he has learned or can learn to adjust himself to his environment." According to Dearborn, intelligence is "the capacity to learn or to profit by experience." Both of these definitions imply that intelligence has something to do with physical, as well as purely conscious, adjustment, but Terman asserts that "an individual is intelligent in proportion as he is able to carry on abstract thinking," while Henmon defends the view that "intelligence is intellect plus knowledge."

For our present purpose, however, a defensible defini-

tion of intelligence is not so essential as an understanding of what is measured by the so-called "intelligence test." Many recent writers have pointed out that such tests do not measure all phases of intelligence. For example, a person may be very adept in music, in mechanics, or in maintaining desirable social relations, yet the ordinary test of intelligence would not reveal this fact. As a matter of fact, intelligence tests make no pretense of measuring such characteristics. What they have been concerned with is the measurement of ability to do school work. Perhaps it would have been fortunate had they been called at the outset by some such name as "tests of ability to do school work," or "academic ability tests." Since the term "intelligence test" has come to be so widely used, however, most tests of this character have come to be known by that name, though the terms "psychological examination" or "test of mental ability" are also in common use.

When does intelligence mature? For practical purposes it is desirable to know the approximate age at which intelligence matures. There are some who look upon intelligence as wholly an inherited characteristic, while others regard it as the product of environmental and hereditary influences. This is not the place to discuss the merits of these contentions, but it does seem appropriate to point out that at least some aspects of intelligence are inherited. Just as the individual matures in his physical equipment, there must also take place a maturing in intellectual equipment. Some investigators who based their conclusions on the evidence obtained from the Army Alpha and Beta tests were inclined to the view that intelligence matured as early as the age of thirteen. Terman had revised the Binet test on the assumption that the age was about sixteen, but later appeared to believe that fifteen might

be nearer the truth. More recent investigations such as those of Chipman¹ and Brown² conclude that fifteen is more nearly correct. Because of the influence of Terman's early work, however, sixteen is still regarded as the average age of maturity by most test makers. The student needs to bear in mind that individual differences are as marked with respect to the maturing of intellectual traits as of physical traits.

The Intelligence Testing Movement

The Work of Binet. Although some suggestion of measuring intelligence appears in earlier educational literature and although some crude attempts at measurement had previously been made by Galton, Cattell, and others, it remained for Binet and Simon to devise the first intelligence scale which was really effective. The test as originally designed was prepared to determine "what pupils should be eliminated from the ordinary school and admitted into a special class." This scale, published in 1905, consisted of thirty exercises arranged in the order of increasing difficulty. In 1908, after considerable experimentation, the scale was revised with grading by years. Again in 1911, just before the death of Binet, the scale was published in revised form, this time with fifty-four exercises graded according to suitability to the child of three years to the adult.

American Revision of the Binet-Simon Scale. The work of Binet and Simon attracted considerable attention in America, and shortly after the death of Binet revisions

¹ Chipman, C. E., Constancy of the Intelligence Quotient of Mental Defectives, *Psychological Clinic* 18:103-111, May 1929.

² Brown, Ralph R., The Time Interval between Test and Re-test in Its Relation to the Constancy of the Intelligence Quotient, *Journal of Educational Psychology* 24:81-96, February 1933.

and adaptations for use in this country were made by Goddard and by Kuhlmann. Somewhat later Terman did an outstanding piece of work in revising and extending the scale and adapting it for use with American children. Still more recently the test has been revised by Terman and Merrill and two forms are available, each one carefully standardized. Each form consists of 128 items arranged in groups for the various age levels from two years to the "superior adult" or twenty-year-old level. The tests include items of various types such as counting, detecting absurdities, disarranged sentences, visual and auditory memory, vocabulary, weight discrimination and judgment.

The original Stanford-Binet test has been regarded by many as the most valid and reliable instrument for measuring academic ability. It was most adequate for children from five to fifteen years of age since above and below these limits the tests were not sufficient in number. The ten years of work which have been put into the careful and thorough revision and restandardization has undoubtedly produced a superior measuring instrument but it is of too recent date for many reports regarding its practical use in the field to have been published. However, Cattell³ reports that for use with pre-school children between the ages of three and six years she has found it "markedly superior to the old forms and to any other intelligence test available for these ages." Neither must it be thought that any of the tests are absolutely infallible for, as was pointed out in an earlier chapter, measurement of mentality has by no means attained the perfection reached by measurement in the natural and physical sciences.

³ Cattell, Psyche, *Intelligence Tests, Review of Educational Research* 8:221-228, June 1938.

Early Group Tests of Intelligence. The Binet-Simon tests can be given by one examiner to only one person at a time. The examiner must, moreover, be well trained for the task of administering the test for much depends upon the skill with which it is done. For these reasons, early workers in the field felt that it would be desirable to devise group tests which could be given by examiners without special training and to groups of individuals at one time. As early as 1917 Otis prepared such a test. Soon after the United States entered the World War a committee from the American Psychological Association set to work on a group test for use in the army. Otis contributed the results of his studies and members of the committee made use of the work of many others as well. The resulting tests, known as the Army Alpha for literate English-speaking soldiers and the Army Beta used with illiterates and soldiers using a foreign language, were administered to about two million men, thus giving a tremendous impetus to the group-testing movement. While these tests did not give as accurate results as might be obtained from individual tests, they did serve a very useful purpose in aiding army officials to select those men who had sufficient mental ability to become officers or non-commissioned officers provided that they also possessed the other essential characteristics of a soldier's equipment. It was also possible to segregate those of very limited mental equipment who would be unable to profit from the instruction afforded by the army. By the use of the tests, this information could be obtained easily and quickly, eliminating delay in assigning a man to that branch of the service for which he was best fitted. These tests were later given to large numbers of college and high-school students until they were supplanted by more suitable tests.

Group Tests for the Elementary School

Despite the limitations of group tests for determining mental ability, some excellent tests of this type have been developed for the elementary school, particularly for grades above the second. While the types of items included in most of these are similar, there are two types of arrangement which may be contrasted by citing examples of each, drawn from earlier tests to appear in this field.

The Haggerty Intelligence Examination,⁴ Delta 2, for grades 3 to 9, consist of six separate parts, for each of which there is a fixed time limit. Exercise 1 is in the form of questions to be answered by Yes or No and is in reality a vocabulary test and a reading test. Exercise 2 is a test of arithmetic problems. Exercise 3 is a test of ability to supply the missing parts of various pictures. Exercise 4 is a same-opposite test. Exercise 5 is called a test of common sense and Exercise 6 is a test of general information.

The second type of arrangement may be illustrated by the Otis Self-Administering Tests of Mental Ability⁵ for grades 5 to 9. In this test the items of various types are miscellaneously arranged so that the pupil must constantly shift from one type to another. The following items from Form A will serve as illustration:

An electric light is to a candle as a motorcycle is to (?)

- (1) bicycle, (2) automobile, (3) wheels, (4) speed,
(5) police ()

Which one of the words below would come first in the dictionary?

- (1) march, (2) ocean, (3) horse, (4) paint, (5)
elbow, (6) night, (7) flown..... ()

⁴ Published by World Book Company

⁵ Published by World Book Company. Quoted by permission.

The daughter of my mother's brother is my (?)

- (1) sister, (2) niece, (3) cousin, (4) aunt, (5) granddaughter ()

One number is wrong in the following series.

What should that number be?

- 3 4 5 4 3 4 5 4 3 5 ()

Which of the five things below is most like these three: boat, horse, train?

- (1) sail, (2) row, (3) motorcycle, (4) move, (5) track ()

If Paul is taller than Herbert and Paul is shorter than Robert, then Robert is (?) Herbert.

- (1) taller than, (2) shorter than, (3) just as tall as, (4) (cannot say which) ()

What is the most important reason that we use clocks?

- (1) to wake us up in the morning, (2) to regulate our daily lives, (3) to help us catch trains, (4) so that children will get to school on time, (5) they are ornamental ()

A coin made by an individual and meant to look like one made by the government is called (?)

- (1) duplicate, (2) counterfeit, (3) imitation, (4) forgery, (5) libel ()

A wire is to electricity as (?) is to gas.

- (1) a flame, (2) a spark, (3) hot, (4) a pipe, (5) a stove ()

If the following words were arranged in order, with what letter would the middle word begin?

- Yard Inch Mile Foot Rod ()

Those who favor the arrangement exemplified by the Haggerty Test contend that the pupil should have a series of similar tasks in order that he may be well adjusted to the tasks. Those who favor the omnibus arrangement, on the other hand, assert that this facility in adjusting one-

self to various new situations is one of the general factors of intelligence which may be tested in this manner.

It will be noted upon examination of tests for intermediate and upper grades that linguistic elements predominate. Items dealing with vocabulary in some form—either definition of words, giving synonyms or antonyms, verbal analogies, supplying the missing word in a sentence, rearranging sentences, etc.—are to be found in practically every group test in which some reading is used and in many individual tests as well. The reason for the inclusion of such items is the fact that items of this sort give the best index to mental ability. Terman, for example, declares that the vocabulary section in the Stanford Revision of the Binet-Simon test gives the best single index of the child's intelligence.

One of the more recent group tests which has found considerable favor is the Kuhlmann-Anderson⁶ test. While this series of tests can be given in grades 1 to 12, it can be given to only one grade at a time, or to only a half grade at a time, since the material used varies for the different grades and in some cases for each half of a grade term. The arrangement is somewhat similar to that of the Haggerty test which employs several sub-tests with sample exercises for each section. A unique feature of the test is the method of finding the mental age equivalent. Instead of referring the pupil's total score to a table of mental age equivalents, the score on each of the sub-tests is referred to a table and the mental age equivalent on each section is determined. The pupil's mental age is then said to be the median of the mental ages found on the various sub-tests. The test appears to be carefully constructed and standardized and mental age equivalents correlate highly with those found by the Stanford-Binet. The

⁶ Published by the Educational Test Bureau.

test would probably not be well-adapted to a schoolroom in which pupils of different grades are found, since a different series of sub-tests would be needed for each grade group.

Another one of the newer tests which has come into very wide use is the Henmon-Nelson Test of Mental Ability,⁷ which has ninety items arranged in the omnibus-cycle form similar to that employed in the Otis test discussed above. The elementary form of this test is designed for grades 3 to 8. Since it is printed in the Clapp-Young Self-Marking series, the pupil's score can be obtained in a remarkably short time. To further facilitate its use, a convenient table has been devised for converting scores into intelligence quotients by referring the pupil's score to his chronological age.

A still more recent test of mental ability, which will find favor among those who are willing to spend considerably more time both in administration and scoring, is the California Test of Mental Maturity.⁸ There is a pre-primary battery, a primary series for grades one to three, an elementary series for grades four to eight, an intermediate series for grades seven to ten, and an advanced battery for grades nine to the adult level. The administration of each test requires at least an hour and thirty minutes and it is recommended that the time be divided into two periods of forty-five to fifty minutes each.

Unique sections of the test comprise sub-tests of visual acuity, auditory acuity, motor co-ordination, memory, and spacial relationships. Scores on the first three of these do not affect the total score but are considered useful indices. Ability to read is not required for the primary test and

⁷ Published by Houghton Mifflin Company.

⁸ Published by the Southern California School Book Depository, Ltd.

the higher tests are less dependent upon reading ability than is true of most group tests at these levels.

Kindergarten and Primary Group Tests

The tests at this level are as a rule composed more largely of pictures and other non-linguistic symbols. Because of the limited reading ability of the pupils, the construction of tests for this level presents difficulties not encountered in tests for the upper levels. Considerable ingenuity has been exhibited by the authors of such tests, however, so that, while they are usually less reliable than those for higher levels, they do give a fairly good indication of a pupil's intelligence.

As an illustration of the types of items often found in these tests, the following sections are found in the Otis Group Intelligence Scale, Primary Examination.

1. Following directions (in marking pictures)
2. Association (of signs and pictures of fruits)
3. Picture completion
4. Maze
5. Picture sequence
6. Similarities (of pictures)
7. Synonyms and antonyms (pronounced to pupils)
8. Common sense (items read to pupils)

The Otis test is designed for the kindergarten and for grades one to four. Other widely used tests at this level, besides the Kuhlmann-Anderson test discussed above, are the Pintner-Cunningham Primary Mental Test for kindergarten and first and second grades, the Kingsbury Primary Group Intelligence Scale, the Detroit Intelligence Tests (separate tests for kindergarten, first grade, advanced first grade, and second to fourth grades) and the Haggerty Intelligence Examination, Delta 1, for grades one, two,

and three. In addition to these, the Merrill-Palmer tests, which must be given individually by a trained examiner, are very useful for children of pre-school and kindergarten age, particularly for children from two to five years. Another individual test at the same level is the Minnesota Pre-School Scale, designed for children of eighteen months to six years of age. Each of these tests requires a considerable amount of testing material which, however, can be procured from the same publishers.

Non-language Intelligence Tests. Non-language tests are not only desirable for use with very young children but also have their place in the testing of older illiterates, foreign born, and the deaf. A good illustration of such a test is the Pintner Non-Language Test, which does not presuppose an English-speaking environment. The test is well adapted to children in grades 3 to 8 inclusive, and consists of six tests as follows:

1. Movement imitation (in which the child reproduces the movements of a pointer after it has been moved from dot to dot in different ways on the black-board).
2. Easy learning (digit symbol type involving the learning of new associations).
3. Hard learning (same type as above).
4. Drawing completion (drawing in the missing parts of pictures).
5. Reversed drawings (reproducing geometric forms as they would be when inverted).
6. Picture reconstruction (indicating by numbers the positions of the parts of pictures so as to make a complete picture).

Other commonly used non-language tests are the so-called form board tests in which the problem is to fit pieces of varying sizes and shapes into the proper aper-

tures. Some of these are so simple as to be useful only with the feeble-minded, whereas others are sufficiently complex to be useful even with superior adults.

Mental Age and Intelligence Quotients

As was mentioned above, the scores on most group intelligence tests may be converted into mental age equivalents by reference to tables. In case of the Terman-Merrill tests the score is found directly in terms of mental age equivalents. For example, a child who is tested passes all of the tests for age six. Of the seven-year tests he passes four of the sub-tests, each of which adds two months to his mental age. In the eight-year group he passes only one test which, likewise, adds two months to his mental age. His mental age is thus the sum of his scores, as follows:

Six-year test	— 6 years, 0 months
Seven-year test	— 8 months
Eight-year test	— 2 months
	<hr/> 6 years, 10 months

The mental age indicates level of intelligence and is designed to be the same as the ability of the average child of the same chronological age. The child above, then, may be said to have attained the same level of ability as is ordinarily attained by a child whose chronological age is 6 years and 10 months.

The intelligence quotient is indicative of the rate of mental growth and is sometimes referred to as the index of brightness. The child mentioned above was tested on his sixth birthday. The intelligence quotient is the ratio of mental age to chronological age and is found by dividing mental age by chronological age and multiplying this quotient by 100. For this purpose it is most convenient

to convert both mental and chronological ages to months, so we have

$$I.Q. = \frac{M.A.}{C.A.} \times 100 = \frac{82}{72} = 1.14 \times 100, \text{ or } 114$$

Using data from Terman and Kuhlmann, Trow⁹ has devised the following table indicating the nomenclature commonly used in designating the various levels of intelligence.

CONVENTIONAL NOMENCLATURE OF DIFFERENT
INTELLIGENCE LEVELS

The last three are in the feeble-minded class; the upper three, because of the undependability of the mental age units, are usually recorded in terms of scores on particular tests.

<i>Adult M.A.</i>	<i>Classification</i>	<i>I.Q.</i>	<i>Per Cent of Population</i>
	genius	175 and above	1
	precocious	150-174	
	very superior	130-149	
(19-21)	superior	120-129	5
17.5-19	bright	110-119	14
14.5-17.5	normal, average	90-109	60
13-14.5	dull, backward	80-89	14
11-13	borderline	70-79	5
8-11	moron	50-69	1
4-8	imbecile	25-49	
1-4	idiot	0-24	

The above classification must be understood as not being altogether definite. It appears, for example, that not all persons with I.Q.'s below 70 are definitely feeble-minded, whereas others with I.Q.'s slightly in excess of that figure might be so classified. The I.Q. generally remains about constant through the individual's life when

⁹ Trow, William Clark, *Educational Psychology*, Houghton Mifflin Company, 1931, p. 128.

accurate measurements are taken. In a few cases where environment has been radically changed, some sizeable differences have been noted after several years, but most of these may probably be attributed to lack of exactness in measurement rather than to actual changes in the individual's intelligence. The teacher should be extremely cautious about considering the results of a single test, particularly a group test, as an absolute indication of intelligence, especially where the pupil rates very low. High scores are probably more significant than very low scores, since the pupil may have had an unfortunate attitude toward the task, an unavoidable interruption in his work, or some other reason for not doing himself justice. Even high scores may within certain limits be due to chance fluctuations, particularly in tests in which chance success plays a considerable part.

School Progress and Intelligence

Although intelligence tests have not yet reached a stage of perfection, there is ample evidence that such tests do give a fairly reliable indication of the child's progress in school. As has been pointed out, it is advisable to use at least two tests, if group tests are used, in order to secure an adequate index. And it may also be pointed out that two individual testings are much better than one. If the results of the two testings are in substantial agreement, one would be justified in prophesying somewhat as Terman does in his *Intelligence of School Children*, where he says in part: ¹⁰

The typical child of 60 or 65 I.Q. tends to remain in the first grade until the age of ten or eleven years,

¹⁰ Terman, L. M., *The Intelligence of School Children*, Houghton Mifflin Company, 1919, pp. 163-164.

and not to reach the fifth grade until the age of fourteen or fifteen years. By this time he has a mental level of only about nine years and is not able to do the school work satisfactorily above the third or fourth grade.

The typical child of 75-79 I.Q. reaches the fifth grade by the age of thirteen years, and if he remains in school is likely to be found in the eighth grade by the age of sixteen or seventeen. Nearly always, however, his grade location is higher than the mental age would warrant.

Children of 80-84 I.Q. usually remain two years in the first grade, and complete the eighth grade, if they complete it at all, one or two years behind schedule time.

On the other hand, children of 120-129 I.Q. are usually found either one or two grades accelerated. Nearly all of this gain, however, is made in the first year or two of school life. After the first year, they are held to the one-grade-one-year pace of average children. Even so, the central tendency is for them to complete the eighth grade at the age of thirteen.

The situation is slightly but not proportionately better for the I.Q. group of 130-139. Children of 140 to 170 I.Q., however, are likely to become three or four years accelerated and to reach the eighth grade by the age of eleven or twelve years. Wherever children of the higher I.Q. groups are located, their work always presents a striking contrast with that of children of the 60, 70, or 80 I.Q. class who are several years their seniors.

These predictions are probably less valid under present conditions than they were when the above was written since the present-day tendency in many schools is to promote all pupils each year regardless of their academic achievement in order to keep them associating with pupils of their own ages and of similar social adjustment.

Some critics of the testing movement cry out against what they consider "determinism," insisting that such classifications as indicated above are not democratic. It is true that an absolute denial of equal educational opportunities to children of low I.Q. would be unfortunate. However, it is the hope of the proponents of intelligence tests that such classification will enable school authorities to place children in learning situations most beneficial to them. There are many things which individuals even of the moron type can be trained to do and every individual should have an opportunity to train for a useful life. A knowledge of the pupil's mental ability makes possible a much more satisfactory, as well as an earlier, adjustment.

Uses of Intelligence Tests

The most common single use to which the results of intelligence tests are put appears to be that of classifying pupils for purposes of instruction. While this practice has been condemned by some, most school executives have come to feel that such homogeneous grouping is desirable, at least in some portions of the school system. Those who argue against such practice put forth several arguments, one of the most common being that heterogeneous grouping is more democratic. Persons who use this argument are usually unaware of the very wide range of individual differences, or ignorant concerning the optimal conditions for learning, or both. Certainly there can be no lack of democracy in a situation which improves every child's opportunity for success. A more detailed discussion of the advantages and disadvantages of homogeneous grouping appears in Chapter XIII.

Grading is quite properly done on the basis of pupil performance in the classroom and on teacher-made tests

of accomplishment or standard achievement tests. Promotions are generally determined in the same way. Yet in many instances data from intelligence tests prove to be very helpful, particularly when considering the advisability of making extra promotions. In some schools grades are assigned in accordance with the pupil's ability; that is, a student of inferior ability may receive a superior grade if he does the type of work for which pupils of superior ability receive only an average grade. More will be said about this practice in a later chapter.

A better understanding and handling of pupils is often obtained from examining intelligence test data. It is evident that school work should be suited to the intelligence of the pupils. Even disciplinary problems are often easier to understand and to solve if test data are available, for while a teacher may pride herself on her ability to estimate quite accurately the ability of each pupil after a relatively short acquaintance, it has been shown that mistakes are frequently made even by the ablest teacher. Furthermore, the teacher is often called upon to present evidence of her opinion of a child's mental ability. Data from an intelligence test furnish her with something tangible and objective with which to argue her case. Pupils may even be motivated to do better work if they understand that the teacher is aware of their ability.

Guidance in the selection of a high-school course or in the selection of a vocation can be carried on much more successfully by a teacher or counselor provided with intelligence test data. Here again these data must be supplemented with information concerning the pupil's interests, his special talents, if any, and his achievement in the various school subjects.

Researches of various sorts can reach definite and defensible conclusions only with the aid of intelligence tests.

For example, the teacher who is interested in knowing which of two teaching methods is better may try one method with one group and the other method with another group and examine the pupils at the close of the period of instruction to secure an answer to her problem. If one group has superior ability, she may easily come to the wrong conclusions, particularly if she is not aware of its superiority. The data from an intelligence test will help her to avoid drawing the wrong conclusion.

Uses to which intelligence test data are put in the schools have been summarized by Deffenbaugh¹¹ in the table on page 269.

For many of these purposes the intelligence test data serve as only a partial guide. For example, in homogeneous grouping many other factors such as achievement test scores, work in class, age, social adjustment, physical qualities, emotional characteristics, personal appearance, and the like play important roles.

PROBLEMS

1. Can you think of anything other than intelligence which we are able to measure and yet cannot define adequately?
2. Discuss the relative merits of the "omnibus-cycle" arrangement of intelligence test items and the sub-test arrangement. Which is more economical of testing time?
3. What is the intelligence quotient of a pupil whose mental age is 6 years, 8 months and whose chronological age is 8 years, 2 months? Of a pupil whose mental age is 11 years, 4 months, and whose chronological age is 9 years, 10 months?
4. Would you inform the parents of either of the above children concerning the intelligence rating? Would you inform the children?

¹¹ Deffenbaugh, W. S., *Use of Intelligence and Achievement Tests in 215 Cities, City School Leaflet*, No. 20 (March 1925). By permission United States Office of Education, Washington, D. C.

<i>Purposes for which tests are used</i>	<i>Elementary Schools</i>		<i>Junior High Schools</i>		<i>High Schools</i>	
	<i>Per cent of cities</i>	<i>Rank of purpose</i>	<i>Per cent of cities</i>	<i>Rank of purpose</i>	<i>Per cent of cities</i>	<i>Rank of purpose</i>
Classification of pupils into homogeneous groups	64	1	56	1	41	1
Supplementing teachers' estimates of pupils' ability . . .	62	2	44	2	33	2
Diagnosis of cause of failure	46	3	29	3	24	3
Establishment of classes for subnormal children	43	4	14	10	7	13
Extra promotions	40	5	21	4	8	11
Comparison with other school systems	26	6	18	7	13	7
Admission to first grade of elementary school	25	7	0	21	0	21
Placement of new pupils from other schools	23	8	19	6	10	10
Regular promotion of pupils	22	9	15	9	6	15
Determining comparative efficiency of teachers	20	10	11	13	10	9
Establishment of classes for super-normal children	20	11	6	17	2	20
Diagnosis of cause of success	19	12	16	8	12	8
Demotions	17	13	8	16	7	12
Determining changes in method of presentation of lessons	14	14	13	11	6	15
Determining changes in subject-matter of courses of study	11	15	9	14	7	14
Determining class marks . . .	10	16	7	19	4	18
Establishing special supervised study groups	8	17	6	17	3	19
Vocational guidance	0	—	13	12	17	6
Determining number of courses to be carried at one time by high school pupils	0	—	9	14	21	5
Guidance in the selection of high school course	0	—	19	5	24	4
Admission to organized school activities	0	—	3	20	5	17

5. What would be the difficulties encountered in sectioning pupils according to mental ability when they enter the first grade? Do you think this procedure would be helpful?
6. What particular aspects of intelligence does the typical intelligence test fail to measure? Are these important?

BIBLIOGRAPHY

- Boynton, Paul L., *Intelligence, Its Manifestations and Measurement*, D. Appleton and Company, 1933.
- Brown, Ralph R., The Time Interval between Test and Retest in Its Relation to the Constancy of the Intelligence Quotient, *Journal of Educational Psychology* 24:81-96, February 1933.
- Chipman, C. E., Constancy of the Intelligence Quotient of Mental Defectives, *Psychological Clinic* 18:103-111, May 1929.
- Deffenbaugh, W. S., Use of Intelligence and Achievement Tests in 215 Cities, *City School Leaflet*, No. 20, March 1925, United States Office of Education, Washington, D. C.
- Douglas, O. B., and Holland, B. F., *Fundamentals of Educational Psychology*, Chapter XIX, The Macmillan Company, 1938.
- Pintner, R., *Intelligence Testing*, Henry Holt and Company, 1931.
- Terman, Lewis M., *The Intelligence of School Children*, Houghton Mifflin Company, 1919.
- Terman, Lewis M., and Merrill, Maude A., *Measuring Intelligence*, Houghton Mifflin Company, 1937.
- Trow, W. C., *Educational Psychology*, Houghton Mifflin Company, 1931.



CRITERIA FOR THE SELECTION OF TESTS

Chapter XII

CRITERIA FOR THE SELECTION OF TESTS

IT IS THE PURPOSE of this chapter to develop in the mind of the reader certain bases for the evaluation of the standard tests now available. Since none of the present tests are perfect measuring instruments, and since in most fields many tests have been devised, some of them well constructed and others poorly constructed, it is necessary for the teacher to be able to select the best test available for the purpose.

The Purpose of Testing. As has been pointed out in previous chapters, standard tests have been used for a variety of purposes. It may also be said that much time, effort, and expense have been wasted because examiners had no definite purpose in mind in giving standard tests. In order to secure the most benefit from the testing program, tests must be selected with reference to the purpose which it is hoped to achieve. If, for example, the test is to serve as a basis for comparing the achievements of a given school system or a given grade with the achievements of pupils throughout the country, the test must be adequately standardized. If, on the other hand, the test is to be used for analysis of pupil difficulties or for detailed diagnosis, then care must be taken to secure a test which is sufficiently analytical or diagnostic. Such a test as the Gates Silent Reading Test, which measures abilities

in four different types of reading, may be considered analytical since it points not to specific reading difficulties but merely to types of reading abilities with which the pupil has trouble. In arithmetic, on the other hand, we have several diagnostic tests from which one can ascertain whether a pupil has difficulty in securing the right answer to a simple combination, like nine plus seven, or whether his difficulty lies in the matter of "carrying." Tests for diagnostic purposes must be very detailed and norms may or may not be provided. In any case, they are by no means as important here as in the case of survey tests, since the purpose is quite different. Tests should indicate the possibilities of use of the results. If the test is diagnostic, ample explanation of how to secure as complete diagnosis as possible and how to use the results thus obtained should be included. If the test is a survey test, there should be an indication of what scores of different levels indicate.

If a test is to be used for individual measurement, the reliability of the test must be high, since a test with low reliability may give an *average* of performance which is fairly accurate, but fail to differentiate properly among the various pupils of a given group. Since tests are often used for determining who shall be promoted, for assigning grades to pupils, and for guidance of pupils, scoring must be sufficiently accurate to convey some meaning.

The scope of the test must also be taken into account. If, for example, one wishes to test in American history covering the colonial period only, one should be careful to avoid a test which is concerned also with the national period.

Validity

Curricular Validity. The most important question to be asked about any test is: Does it really measure what we

wish to measure; in other words, is the test a valid one? Usually this is a matter which the test user must decide for himself, though there are some aids contained in the manuals of most of the really good standardized tests.

A test maker may resort to a number of devices in order to insure the validity of a test. He may, for example, assure himself that the items of the test are all contained in the more widely used textbooks. This was the method employed by Denny and Nelson in the construction of their tests in American history, and by Ruch and Popenoe in the construction of their tests in general science. Other test makers select only those items which appear in well-constructed and widely used courses of study. This latter method has the advantage that where curriculum makers have included important subject matter not found in textbooks, such material can also be tested. Since much of the teaching in the elementary schools of the United States is done with rather close adherence to the available textbooks, the results of these two methods are usually very much the same. Some test makers have selected their items by choosing materials used in tests constructed by teachers over a period of years. Tests are also constructed with reference to the recommendations of various national committees, or to the individual opinions of experts in the various fields.

The fact that a test covers only material dealt with in numerous textbooks, courses of study, or other sources, does not insure its maximum usefulness. A history test might, for example, be perfectly valid from this point of view and yet, because it deals only or largely with military events, be unsuited to the needs of a teacher who has stressed the social, economic, and political phases of the history of the same period. Frequently, the only way of discovering such a fact is by careful reading of the in-

dividual test items. Again, the teacher might discover that a certain test on the national period of American history was so largely concerned with the Civil War period that the test would not be suited to a course of study giving fairly uniform attention to the entire period from 1789 to the present time. Again, despite the fact that the items in a test may have been validated by reference to many textbooks, it is always possible that for the user of a given text some of the items may not be pertinent. An occasional irrelevant item need not be considered sufficient reason for discarding a test, but a test containing many such items should be rejected.

Some achievement tests are so constructed that it is doubtful whether they really test in the field for which they are designed. A so-called geography test may, for example, be in reality a test of reading ability. Likewise, a test designed to measure health knowledge might, in fact, be testing pupils' understanding of vocabulary.

The test user must be on the alert to see that the items meet his demands in every way. They should have suitable vocabulary, should not be catch questions, and, as a rule, should be arranged in order of increasing difficulty. This last factor not only affects the validity of a test, but also to a greater degree its reliability.

Statistical Validity. Test makers do not always content themselves with the above methods but sometimes resort to other methods of ascertaining that they really are valid measuring instruments. One of these methods is by correlating the test scores with some criterion of known validity. Let us suppose, for example, that one wished to construct a test of speed in brick-laying. It would be a relatively simple matter to administer the test to a number of brick-layers and to count the number of bricks which each is able to lay per hour or per day. If the scores on

the test were correlated with the number of bricks laid in a unit of time, the resulting coefficient would indicate whether or not the test was really a valid test of speed in brick-laying.

In most educational tests a criterion of such known validity is not ordinarily to be had. As a result, test makers often content themselves with correlating test scores with grades assigned by teachers of the subject in question. Since teachers' marks are notoriously unreliable, it is not to be expected that correlations between test scores and grades will be high. About all one can hope for in such a case is a reliability coefficient of about .60 or .70.

There is still a third criterion used in determining the validity of a test and that is the scores on another test of known validity. The Stanford Revision of the Binet-Simon scale is almost universally regarded as a valid measure of intelligence or mental ability. If, then, the author of another test of intelligence were to find that the results on his test correlated very highly with those obtained from the Stanford Revision test, he could rest assured that his test was valid.

Mechanics of Administration

Most standardized tests furnish rather detailed and explicit directions for administration. This is very important for unless the procedure used in administering a test is the same as that used when the norms were established, the results are not comparable. The exact words supplied in the manual of directions should be used in the administration of a standardized test, but it is best for the examiner to so familiarize himself with the directions that they can be given in a natural way and with only occasional reference to the printed directions. Occasionally

the administration of a test requires considerable training, but this is not true for most group tests. Tests with very complicated directions will be avoided by the teacher who is not familiar with testing techniques. As an illustration of things to look for in examining a test from the standpoint of mechanics of administration, the following questions may be raised:

1. Is the pupil told what to do when a page is completed? Uncertainty here may lead to a waste of valuable time.

2. Is the pupil told what to do if he finishes before time is called? In some cases the pupil is permitted to go back over his work, while in other cases this is specifically forbidden. If some pupils who finish early do go back while others refrain from so doing, the results are not strictly comparable.

3. Are there sufficient fore-exercises so the pupil understands how his answers are to be indicated, and the general nature of the items?

4. Are the time limits adequate for sufficiently accurate testing? If, for example, one desired to secure a rather accurate measure of reading ability, he could be quite certain that a four- or five-minute test would not suffice.

5. Is the test short enough so that it can be administered in the period available? If, for example, the work is departmentalized with periods of forty minutes, one would want to be sure that the time required for taking the test, plus the time for distributing papers, giving directions, etc., was not in excess of such a period.

Mechanics of Scoring

While it is true that the time required for scoring tests, tabulating data, interpreting results, and the like is not all-important, it does present one of the problems which is all too often overlooked in planning a testing program.

Needless to say, a test should be thoroughly objective. The importance of objectivity is seen not only in the fact that results may not be comparable if much of the scoring is subjective, but also in the fact that a great deal of unnecessary time may be consumed if a scorer must weigh the worth of various answers. The matter of time for scoring is more important than may appear at first glance; the time required per paper may seem insignificant until one multiplies it by the number of pupils to be tested. Let us suppose, for sake of illustration, that two equally good tests are available to a teacher of forty pupils. One of the tests requires six minutes of scoring time; the other requires but a half minute. To score the first test would require four hours, or three hours and forty minutes more than would be required for the second test. When one considers how busy the typical teacher is and how valuable this additional time might be if spent in diagnosis of difficulties and in planning remedial instruction instead of in clerical routine, the advantage of the test which can be scored quickly becomes apparent. In a few schools teachers are provided with clerical assistance, somewhat minimizing the advantage of a test which can be scored quickly. When clerks are used, the scoring should be so mechanical that it can be done by untrained persons.

In some of the older tests the various items were assigned different values in accordance with their relative difficulty. For example, in the first edition of Monroe's Silent Reading Test for Grades 6, 7, and 8, item number 1 is given a comprehension value of 2; item number 4, a comprehension value of 3; and so on until the final paragraphs have a comprehension value of 5. The theory underlying this procedure was that a pupil who was able to do a more difficult task should be allowed more credit than a pupil who was able to do only a relatively simple

task. Such weighting complicated the scoring and has been almost entirely abandoned since the discovery that the number of correct items on a test and the sum of these scale values correlated well over .90, or, in other words, gave almost the same index of a pupil's ability. It is, however, desirable that the items in a test be arranged in order of increasing difficulty, in order to allow those pupils who do poorly to work at tasks within their capacity. No test should be so difficult that some pupils will make no score, nor so easy that many pupils will make perfect scores.

The General Make-Up of the Test

Test publishers have usually taken great care to print tests legibly and on a satisfactory quality of paper. Illegibility may result from print too small for younger children, but until we have more conclusive studies on the size of type children of various ages are able to read with facility, no definite standards for size of type can be set up.

For the most part, tests are available in the quantities desired, but it sometimes happens that only lots of one hundred or more are available. The teacher who plans to test a smaller group will, therefore, wish to inform herself on this point. Since most tests will be kept on file for some time, it is well to note whether the size of the test booklet permits convenient filing.

Most of the best standardized tests have at least two duplicate forms and sometimes more. The teacher will often wish to use a second form at a later date in order to measure pupil progress. She may also wish to secure a more reliable measure than can be obtained by a single form, in which case the alternative form provides an easy way of increasing reliability through added length. One

might be tempted to say that the more forms a test has, the better. However, in many fields the amount of good, valid material is soon exhausted so that the maker of several alternative forms must resort to including less valid and less significant items. This will not be true in the case of tests of reading or intelligence, in which fields there appears to be an inexhaustible supply of material, but might easily be true in such fields as geography, health, or history.

In a few tests, because of the conspicuous manner in which answers are indicated, the pupil who wishes to profit from the efforts of his neighbor may do so with comparative ease. The teacher who is desirous of obtaining a valid measure will avoid such tests, or, if this is not feasible, will seat the pupils in such a way that this danger is minimized.

Another practical consideration in the choice of a test is its cost. While other considerations are of more importance, school budgets generally have very definite limitations. The most expensive test is not always the best; in fact, there appears to be no relationship whatsoever between the cost and the merits of tests. It also happens that in some cases the first cost may be relatively large, but because all or portions of the test materials may be used repeatedly, in the long run they are cheaper than those whose first cost is low.

Norms

In the selection of a test for comparative purposes, the user will do well to assure himself that norms are provided and expressed in the most useful way. As a matter of fact, standard tests may be distinguished from informal objective examinations mainly by the fact that they do

have norms. It is true that the standard test, as a rule, is very carefully constructed, usually goes through several experimental processes, and the author is likely to make sure of the validity, scaling, and other features of his test. Yet all of these things might be true of an informal test; hence, it is the fact that a test has norms which distinguish it from an informal or classroom test.

In the establishment of norms, test makers attempt to have the tests given to large numbers of pupils who are typical of the group for which the test is designed. For example, if one wished to standardize a fifth grade test, one would have the test given in schools in which the fifth grade group was not distinctly superior or distinctly inferior in the particular field in question. In order to guard against such "selection," the author would probably choose some rural schools, some small town schools, and some schools in cities of varying size. In the cities he would guard against selecting all of the schools from the better residential districts or all of them from the factory districts, etc. The user of the test could then be informed concerning performance under ordinary or "average" schoolroom conditions. Some authorities have advocated separate reports of norms for various types of schools so that a teacher in the school of a well-to-do residential district of a large city could compare the performance of her pupils with that of the pupils in other similarly situated schools. Aside from the practical difficulties involved in securing representative samplings for each group, such a procedure implies that pupils in the various types of schools vary greatly. Investigation has revealed some such differences, but a good school in a poor city district often shows better achievement in some school subjects than does a mediocre school in a better residential district. The writer is, for example, familiar with one elementary

school in a city system which invariably has lower median scores on all tests than the other schools in the city, except only in the field of arithmetic computation in which this school takes the lead. In other words, there is so much overlapping among the various types of schools that the practice of providing separate norms has not come into common practice.

A norm represents the performance of the typical individual on the test in question. Thus, about as many pupils exceed the norm as fall below it.

The teacher should avoid thinking of the norm as the desired standard which her pupils should attain. Probably the standard or goal towards which the teacher should strive, if expressed in terms of the score on a test, should be somewhere above the norm. However, in some instances the pupils of a given room may be so far below a given norm that the teacher will want to set up as her immediate objective a point not as high as the norm, but a point somewhere below it.

There are various ways of expressing norms of performance in the elementary school and each of these methods has some advantages. The most commonly designated norms are "grade norms," "age norms," and "percentile norms," while T-scores and other forms of norms have also been employed by some test makers.

Grade Norms. Norms for the various school grades and fractional parts of grades are among the most commonly reported and frequently used. These norms are ordinarily determined by testing a large number of pupils in each of the grades for which the test is designed and determining the average performance. Let us suppose, for example, that on a given test administered at the close of the school year, the third grade made an average of 42 points, and the fourth grade an average of 52. Since the pupils would

have been in their present grade for nine months, we would say that 42 is the equivalent of a grade placement of third grade—9 months. For convenience, this figure is ordinarily written 3.9, the whole number referring to the grade and the decimal to the number of months in that grade. We may now interpolate between 3.9 and 4.9 and give grade values to all scores between 42 and 52 as follows:

Score	42	43	44	45	46	47	48	49	50	51	52
Grade	3.9	4.0	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9

This procedure gives one the impression, which all too frequently is not true, that the test used is a very accurate and highly refined measuring instrument. If one makes allowance for the unreliability of the test, however, the assignment of grade equivalents to test scores is both convenient and desirable, since the grade concept is so readily understood. As one proceeds to the upper grades its usefulness is gradually diminished because of the fact that the amount of retardation, elimination, and acceleration varies greatly in different schools.

Age Norms. Another convenient method of indicating the meaning of scores is that of converting them into age equivalents. The procedure here is much the same as that used in determining grade norms except that in this case all of the scores made by pupils of a given *age* are averaged. Interpolation for the various months can be made as indicated above.

Age norms are particularly useful if it is desired to compare achievement test results with results on tests of intelligence, since practically all intelligence test scores are converted into mental age equivalents. Age norms also permit, of course, comparison with chronological age. Another advantage of age norms is that age is more constant than

grade placement in the various school systems. The teacher does not often need to make a choice between a test with grade norms and one with age norms, since for most of the better elementary school tests both types of norms are reported.

Percentile Scores. Some test makers report percentile scores either in addition to age and grade equivalents or instead of such equivalents. A common form of reporting such scores is illustrated below in the table taken from the manual for the Denny-Nelson American History Test, indicating the important percentile scores for pupils of grade eight who have had one semester and one year of work on the national period of American history.

NORMS FOR THE NATIONAL PERIOD TEST

Grade 8

	<i>One Semester</i>	<i>One Year</i>
1% equal or exceed	87	91
10% equal or exceed	75	82
20% equal or exceed	67	75
25% equal or exceed	65	72
30% equal or exceed	61	70
40% equal or exceed	56	66
50% equal or exceed	52	61
60% equal or exceed	46	54
70% equal or exceed	40	48
75% equal or exceed	38	45
80% equal or exceed	35	42
90% equal or exceed	30	36
99% equal or exceed	18	23

The method of computing percentile scores is practically the same as that employed in computing the median, the essential difference being, of course, in the fractional part of the frequency distribution which is used. To find the 20th percentile, one would count up one-fifth of the cases; for the 40th percentile, two fifths, etc.

Percentile scores have their widest application in high-school and college tests where age and grade norms have little meaning, and in computing data from informal objective tests where no norms have been established.

T-Scores. McCall has introduced what is known as the T-score (in honor of Thorndike and Terman) which has been used in some tests, notably the Thorndike-McCall Reading Scales. It is based upon the standard deviation as the measure of variability and is determined by means of the following formula.

$$\text{T-Score} = 50 + \frac{10 (X - \text{A.M.})}{\text{S.D.}}$$

In this formula, 10 is a constant introduced to eliminate decimals; 50, a constant introduced so that the average T-score will be 50. X is any raw score on the test; A.M. is the arithmetic mean, and S.D. is the standard deviation of the scores.

The advantage of the T-score is that it can be computed for any test in such a manner that a score of a given size on one test is equivalent to a score of the same size on another test, regardless of the length or difficulty. Perhaps the chief reasons that they are not more commonly employed are that they are somewhat difficult for the average person to understand and that they cannot conveniently be compared with norms expressed in terms of grade or age levels.

Adequacy of Norms. The question is frequently raised as to when a test has adequate norms. Should the test be considered adequately standardized when it has been given to 500 pupils or only when it has been administered to 50,000? In general, there has grown up a conviction that if a test has been administered to 500 pupils in each of the grades for which it is designed, then the standardi-

zation is likely to be adequate. However, the number of cases involved is not of such great importance as is the adequacy of the sampling. If one could be sure that the group tested was exactly typical of all pupils of a given age or grade, then a very small number would suffice. Since one cannot ordinarily be certain of the representativeness of the sampling, larger numbers are usually sought. The reason for thinking that 500 cases are ordinarily sufficient for each grade is the fact that experience has shown that with the tabulation of additional data only very minute changes, if any, are made in the norms.

Reliability. In selecting a test the teacher will wish to know whether or not it is a dependable instrument in the sense that the results obtained on one day are consistent with those which would be obtained at any other time under the same conditions. To illustrate from another field of measurement, let us suppose that a teacher were to ask four pupils to measure the length of her desk. One of them reports 58½ inches; another, 60 inches; the third, 59 inches; and the fourth 59½ inches. As she watches the pupils at work she decides that the fault does not lie so much with the children as it does with the linen tape measure which stretches considerably when held taut. When a steel tape is used, the variation in results is much less. We may say, then, that the linen tape is not a reliable measuring instrument, since the results obtained from it are not consistent with themselves.

The reliability of tests is usually reported in terms of the coefficient of correlation, which is discussed in Chapter IV.

In applying the correlation technique to determine reliability of tests, three methods are used, as follows:

1. Correlating scores made on one form of the test with those made on a second form. Since most

standard tests have two forms, this is a most convenient and in many respects the most satisfactory method. The two forms should be given on successive days or with a comparatively short interval of time intervening.

2. Correlating chance halves of the test against one another and estimating the probable reliability of the entire test by means of the Spearman-Brown formula. When this method is used, the score on the even-numbered items is usually correlated with the score on the odd-numbered items. The formula may be written

$$r_{nn} = \frac{Nr_{12}}{1 + (N - 1)r_{12}}$$

where r_{nn} is the reliability of the whole test; N is the number of times which the whole test is longer than the parts and r_{12} is the correlation between the parts (as between odds and evens). Let us suppose, for example, that the correlation found between evens and odds is .70 and that we wish to find the probable reliability of the entire test. Substituting in the formula we have

$$r_{nn} = \frac{2 \times .70}{1 + .70} = .82 +$$

3. By correlating the scores made on the test when the same test is given after an interval of only one or two days and with no coaching intervening. Even without coaching we expect pupils to do somewhat better on the second trial since some pupils will look up items missed, or by conversation with other pupils will be set right on some points, or by considering an item after the test is complete will see an angle which was previously not apparent. A better acquaintance with the mechanics of the test is also an aid to the pupil. If all pupils profited equally, the correlation would not be disturbed, but since they do not, this method is not as satisfactory as either of the above.

The Probable Error of a Score. Since the coefficient of reliability is difficult to interpret and since it is sometimes misleading, some test makers prefer to report the reliability of their test in terms of the probable error of a score. If, for example, the coefficient of reliability for one test is reported as .87 and is based on scores by pupils of the fourth grade alone, it is quite likely that it is a more reliable test than one for which the reliability coefficient is .90 based on pupils in grades three to eight. When the range of talent is decreased, the coefficient tends to be smaller; when it is increased, the coefficient is likely to rise. The probable error of a score takes into account both the variation in the group (as indicated by the standard deviation) and the correlation between scores.

By the probable error of a score is meant the *probability* that a score of a given size would vary by a given amount or, stated in the opposite manner, that it would not vary by more than a given amount. Let us suppose that on a given test a fifth grade pupil made a score of 73. The author informs us that the probable error of a score is 2 points. We may then say that the chances are even that if this pupil were to take a second form of the test, his score would lie between 71 and 75 (73 ± 2). The chances are also even, of course, that this pupil would make a score outside of this range, that is, somewhere below 71 or above 75. However, the chances are almost negligible that the pupil would make a score which varies more than three times the probable error from the score made. In the above case, then, the pupil is almost certain to make a score between 67 and 79.

The value of the probable error concept may be illustrated from the following data from two tests in the same field.

Test I—Fifth grade norm	= 73
Sixth grade norm	= 75
Probable error of a score	= 2
Reliability coefficient	= .90
Test II—Fifth grade norm	= 61
Sixth grade norm	= 74
Probable error of a score	= 3
Reliability coefficient	= .85

In Test I, a pupil who scores 73 (the fifth grade norm) has a good chance of scoring as high as the sixth grade norm if the test is repeated. In Test II the chances are negligible that a pupil who scores 61 (the fifth grade norm) would be capable of reaching a score equivalent to the sixth grade norm.

Factors Affecting Reliability of a Test

There are many influences which have their effect on the reliability of a test. Some of these are in the test itself and others are factors beyond control of the test maker. It is obvious that an objective test is more reliable than one lacking in objectivity since consistency in scoring items which are not thoroughly objective will vary from item to item or from test to test. Since most standard tests are composed of recognition items which can be scored with perfect objectivity, this factor does not enter so often.

The length of a test is a factor in its reliability, since it is known that a longer test is more reliable than a shorter test of a similar nature. The fact that there is a known mathematical relationship between the length of a test and its reliability makes possible the prediction of reliability by means of the Spearman-Brown formula as indicated above.

A test which is evenly scaled is likely to be more reliable than one which is not scaled. By scaling is meant arranging items in order of increasing difficulty and by equal amounts so that, for example, item 6 is as much more difficult than item 5 as item 5 is more difficult than item 4. The term "scaling" is probably more often used in connection with the arrangement of items in accordance with difficulty, as would be represented by the normal curve.

The factors mentioned above refer to the inherent qualities of a test. But there are also conditions in the pupils which affect reliability. Among these may be mentioned fatigue, momentary inattention to the task at hand, enthusiasm (or lack of it), illness, physical injury, etc. While these factors influence test results rather less than is commonly believed, they no doubt have some effect. The familiarity of pupils with the techniques of testing also has some influence as does, of course, the attention given to the administration of a test and to the elimination of extraneous factors.

Test-Rating Scales. A few attempts have been made to devise test-rating scales which would make it possible to express in quantitative terms which of two or more tests under consideration has the more points in its favor. The use of such devices has not become widespread, partly because it is not always desirable to choose a test simply because it rates high on one of these scales. For example, a test might rate very high and yet be lacking in certain diagnostic features very much desired. The chief value of such devices is rather that they do call attention to certain desirable characteristics of tests. Teachers who are not thoroughly familiar with the criteria for the selection of tests will do well to consult such scales. What is perhaps

the most ambitious of such scales is one suggested by Cole and von Borgrsrode ¹ and reproduced below.

SCALE FOR RATING STANDARDIZED TESTS

I. Preliminary Information

1. Exact name of test
2. Name and position of author
3. Name of publisher and nearest address
4. Cost
5. Date of copyright
6. Purpose of test

II. Validity (25)

A. Curricular (15)

1. Exact field or range of educational functions which test measures?
2. Ages and grades for which intended?
3. Criteria with which material was correlated?
4. Do questions parallel good teaching procedures?
5. How wide is sampling of important topics?
6. What is the social utility of questions?
7. Is test claimed to be diagnostic? (If so, proof and see VI, 5, c, below.)

B. Statistical (10)

1. Correlated against what outside criteria?
2. Size of coefficient of correlation?
3. Size and representativeness of sampling?
4. Proof of validity of items (such as statements as to experimental tryout of items individually to determine that no large percentage is failed or passed by all pupils and that the items show a consistent increase of percentages of successes with successive age or grade levels).

¹ Cole, Robert D., and von Borgrsrode, A Scale for Rating Standardized Tests, *School of Education Record of the University of North Dakota* 14:11-15, 1928.

III. Reliability (25)

A. Most important items

1. Correlated with what?
2. Size and representativeness of sampling?
3. Reliability coefficient
4. The means of the distributions
5. The standard deviations of the distributions
6. If some other measure than the above three is given to prove reliability, what is it?
7. Intercorrelations

B. Less important but desirable

1. Order of giving various forms of test
2. Is test reliable enough statistically for individual measurement, or can it be used only for groups?
3. Evenness of scaling (see II, B, 4)
4. Are pupils accustomed to this type of test?

IV. Ease of Administration (15)

1. Manual of directions (3)

- a. How complete and simple is the manual?
- b. Does manual control test conditions well?
- c. Typographic make-up

2. Simplicity of administration (8)

- a. Amount of explanation needed for pupils by examiner?
- b. Are directions to pupils clear, detailed, comprehensive?
- c. Is arrangement of test convenient for pupils?
- d. Are samples and "fore-exercises" given when needed?

3. Alternate forms (3)

- a. Number.
- b. Evidence of reliability
- c. Evidence of equivalency

4. Time needed for giving

V. Ease of Scoring (10)

1. Degree of objectivity—purely objective or some judgment on part of examiner?
2. Are adequate directions given—clear, equal to all emergencies?
3. Is scoring key adjusted to size of test?
4. Time needed to score one test
5. Simplicity of procedure
 - a. Number of processes needed to get final score?

VI. Ease of Interpretation (20)

1. Norms (6)
 - a. Kind—age, grade, percentile, etc.
 - b. Derivation—size and representativeness of sampling
 - c. Tentative, arbitrary, or experimental?
 - d. For separate parts?
 - e. How expressed?
2. Is class record provided?
3. Are there provisions for graphing results?
4. Is interpretation of raw scores easy or hard?
5. Application of results (10)
 - a. Are directions or suggestions given for application of results to benefit teaching or administration?
 - b. Are tests survey or diagnostic?
 - c. If diagnostic—
 - (1) Proof of diagnostic value?
 - (2) What principle or principles underlie construction?
 - (3) How many different skills, abilities, or aspects of the subject are analyzed or measured?
 - (4) Does the analysis of total subjects into unit abilities follow teaching practices or needs?
 - (5) Is the diagnosis individual or class—proof?

- (6) Does the test demand tabulations of individual pupils' errors to secure a diagnosis?
- (7) Is a remedial program provided or suggested?

VII. Miscellaneous (5)

- 1. Typography and make-up
 - a. Arrangement of printed matter
 - b. Legibility of type
 - c. Quality of paper
 - d. Are test blanks free from distractions, norms, directions to examiner, etc.?
- 2. Is the time required for giving as small as is consistent with reliable measurement?
- 3. Is the cost in keeping with the amount, scope, and reliability of the results yielded?
- 4. Is good test service provided by the publisher?
- 5. Kind of new-type questions used?

PROBLEMS

- 1. What is the T -Score of a pupil whose raw score on a given test is 40, if the mean of the scores is 61 and the standard deviation is 7?
- 2. The correlation between the odd and even numbered items on a given test is $r = .76$. What is the reliability of the test?
- 3. A teacher has constructed a test of 250 items in social science and has found its reliability to be represented by a coefficient of reliability of $r = .96$. What is the estimated reliability of 125 items of the same test?
- 4. Why can standard tests ordinarily be scored more accurately and objectively than teacher-made tests?
- 5. Secure at least three complete standardized tests in the same field and rate them on the Cole-von Bengersrode Scale. Would you prefer to use the test which you have given the highest rating? Why or why not?

BIBLIOGRAPHY

- Greene, H. A., and Jorgensen, A. N., *The Use and Interpretation of Elementary School Tests*, Chapter VI, Longmans, Green and Company, 1936.
- Kelley, Truman Lee, *Interpretation of Educational Measurements*, Chapter IX, World Book Company, 1927.
- Odell, C. W., *Traditional Examinations and New-Type Tests*, Chapter II, The Century Company, 1928.
- Smith, H. L., and Wright, W. W., *Tests and Measurements*, Chapter III, Silver Burdett Company, 1928.
- Tiegs, E. W., *Tests and Measurements for Teachers*, Chapter XVI, Houghton Mifflin Company, 1931.

USING EDUCATIONAL TESTS

Chapter XIII

USING EDUCATIONAL TESTS

BECAUSE THE TESTING MOVEMENT, like other new ventures in the field of education, does not now receive as much attention in educational literature as it did a decade or two ago, there are those who believe that there is very little use of standard tests and less emphasis on objective measurement. One who has access to the sales reports of test publishers will discover, on the contrary, that standard tests of both achievement and intelligence are used in very large numbers in all sections of the country and in all types of schools. It becomes evident, therefore, that the reason for the decline in literature on the subject is that tests have come to have an accepted place in schoolroom practices and that controversy is diminishing.

Despite the wide use of tests, there are still those who argue against them for one reason or another. With respect to intelligence tests, for example, there are those who deplore their use because they do not test all of intelligence, because they feel that it is unjust to "brand" a boy or girl with an I.Q. which may be erroneous, or because they as teachers can tell more about the intelligence of their pupils than any paper and pencil test will show. Certainly, as was pointed out in a previous chapter, the intelligence tests do not measure all of intelligence and their authors would be the first to admit it. It remains

true, however, that they do give an index of that phase of intelligence which we call academic ability and that this is a rather important ability, though by no means the pupil's only valuable asset. No one who is familiar with tests and their limitations would wish to "brand" a pupil with an I.Q. even if he felt the I.Q. rating were a very reliable one. There is a vast difference between "branding" the pupil and learning all one can about his abilities or lack of them. Anyone who says that he can judge a pupil's intelligence much better than an intelligence test can determine it shows his ignorance of the many studies indicating that what passes in the teacher's judgment for intelligence is frequently merely a pretty face, a sweet smile, a well-poised bearing, or a well-oiled tongue.

Opponents of achievement tests, and particularly of objective type tests, are sometimes heard to argue that they take too much time which might be spent to better advantage in additional teaching; that they do not test the child's will or determination; that they test merely for information, or that they do not test the whole child. Concerning the first objection, one might raise the very legitimate question as to which teacher is more effective, the one who attempts first to learn what the pupil already knows or the one who goes right ahead with what he terms teaching procedures without any reference to the pupil's background. Achievement tests certainly do not measure determination unless it be in a very indirect way. New tests will have to be built for such traits and it is along these lines that pioneer work in testing is now being done. It is true that the makers of standardized tests have devoted most of their time to testing for information but much progress has been made and is being made in the direction of measuring the pupils' abilities to use their

information in solving problems and arriving at valid conclusions.

No less indefensible than the attitudes indicated above is the apparent feeling of a few over-zealous testers that once some tests have been given all of the problems have been solved, an adequate basis for grading has been secured and one may go on to the next testing period without a backward glance and without further attention to the test results. In the hands of a capable builder the footrule, the level, the plumb line and the carpenter's square are useful devices and one would scarcely award a contract to one who could not use them. Neither would one employ a builder who had no other tools with which to work.

Use of the Pre-Test

Many teachers like to open the work of the school year, particularly in a new position, with a testing program. Such a procedure can easily be justified since through it the teacher may attain a better understanding of the pupils in her group. By using comparable tests at the close of the year she may note the progress her pupils have made. Such a testing program might well include a test of intelligence as well as a standardized test in the subjects taught. Before going ahead with a program of this sort, however, the teacher should acquaint herself with the data already available. Such data may give her the desired information without the use of additional time, effort, and money. She may, for example, discover that the children were given tests of intelligence during the latter part of the previous year. If so, she will probably decide against the administration of another test of this type, for changes in rating are likely to be very slight.

Warning should probably be given concerning the blan-

ket acceptance of achievement test data from the previous spring since marked changes may have taken place during the summer vacation period. These changes are likely to be very noticeable in such drill subjects as arithmetic in which there is often a decided deterioration in the skills developed.¹ In reading, on the other hand, there may be some gains noted, particularly in the upper grades if the pupils have been doing considerable vacation reading.

Another legitimate and helpful use of a pre-test is found in testing prior to entering upon a unit of work in a particular subject. In view of the small number of published tests which are concerned with such limited areas of subject matter, the teacher will ordinarily have to construct such tests for her own use. A careful scrutiny of the pupils' responses will reveal which portions of the proposed unit need most emphasis, and which concepts or items of information are already well known. She may even discover that the children already possess all of the information or the skills she had intended to develop in which case she is free to spend her time in fostering attitudes associated with the material or to proceed at once to another unit of work.

Tests for Grouping

In schools where pupils are sectioned for instructional purposes the question is sometimes raised concerning the most effective tests to use for this sectioning. Should one use an intelligence test or a test in the subject to be taught? For example, if the sixth grade is to be sectioned for instruction in arithmetic, should a test of general men-

¹ Nelson, M. J., How Much Time Is Required in the Fall for Pupils of the Elementary School to Reach Again the Spring Level of Achievement? *Journal of Educational Research* 18:305-8, November 1928.

tal ability be used, or would it be better to use an arithmetic test? Experience has shown that a test in arithmetic is better since sectioning on the basis of general ability results in a less homogeneous group than is secured by the use of an achievement test. If the school is large enough, however, so that there will be several sections in each grade, it is advisable to take into account both general ability and achievement in arithmetic since pupils of low general ability are not likely to proceed as rapidly as those whose general mental ability is high. If sectioning is done in all of the school subjects, however, there will probably develop a very real administrative problem, particularly in the smaller school where the same pupil may belong in a low section in arithmetic, a high section in reading, and a middle section in history. The exact basis for dividing into sections will, therefore, be determined in part by administrative feasibility. For grouping first grade entrants for instruction, either a reading-readiness test or a test of mental ability, or both, will be found helpful.

Among the indices which have been used for sectioning pupils in the same grade for purposes of instruction, the following are fairly common:

1. Achievement as measured by standardized or teacher-made objective tests
2. The Intelligence Quotient
3. Achievement as indicated by school marks of the previous year
4. Mental age
5. Teachers' estimate of ability
6. Some index of social maturity
7. Chronological age

While some attention has previously been paid to the desirability of grouping pupils within a given grade for instructional purposes, it may not be amiss at this point

to enumerate some of the more common arguments pro and con. Those who advocate the procedure usually present the following arguments: ²

1. Grouping improves teaching. If pupils of about the same ability are brought together they will stimulate one another more in class discussion. The teacher can also use her time to better advantage if the group is rather homogeneous.

2. Grouping aids development of social and mental characteristics. The association of people of like interests contributes to social adjustment.

3. Grouping lessens the cost of instruction because the teacher can handle larger groups as effectively as smaller groups of unclassified children.

4. Grouping eliminates many failures. Failures bring discouragement, lessen the confidence of the individual and weaken the morale of the entire school.

5. Grouping aids in the development of initiative and leadership.

6. Grouping makes it difficult for the bright pupil to loaf or to mark time.

7. Grouping keeps pupils interested because they are able to accomplish work which is adjusted to their own level.

Frequent arguments heard from those who are opposed to grouping include:

1. Unclassified pupils duplicate conditions of the ordinary community more nearly than do classified groups since in everyday life all are free to mingle.

2. Unclassified group discussions are more helpful than group discussions in homogeneous groups. Those who have less ability profit from hearing the discussions of the more able; while the more able, in turn, secure valuable practice in leading discussions.

3. Grouping involves administrative difficulties of a

² From Riebe, H. A., Nelson, M. J., and Kittrell, C. A., *The Classroom*, pp. 239-240, The Cordon Company, 1938.

very perplexing sort because adjustments must be made very frequently. If they are not made frequently, the group which started out on about the same level becomes almost as heterogeneous as unclassified groups within a relatively short time. Grouping is also very difficult for small schools.

4. Grouping tends to breed a caste system. Pupils become conscious of the type of grouping employed and the inferior become keenly aware of their inferiority. The brighter ones develop a superiority complex and have such high regard for their ability that they cease to work as effectively as might otherwise be the case.

Tests for Diagnosis

Considerable attention has already been given to diagnostic testing in the various school subjects. Tests of this sort ordinarily cover more minute details than do survey tests, and frequently make it possible for the teacher to note the real reason for incorrect responses. Since published tests of this nature are not available in sufficient number, the teacher should become adept at their construction and should give particular attention to the diagnostic possibilities of all achievement tests. Thus a survey test in spelling may, if the words are properly chosen, reveal the particular study needs in spelling. To make full use of such a test the teacher must not be content with scoring and comparing scores with norms or with comparing the score of one pupil with the median for the group. Rather, she must study each individual error to note where the mistake occurred in each word and the nature of the error.

In preparing a diagnostic test, particularly if the items used permit of chance successes as in the case of most objective types of items, the teacher should see that the same information is called for more than once in somewhat dif-

ferent fashion. A spelling word will serve as a simple illustration. When the word "receive" is pronounced, many children will know that it is properly spelled either "receive" or "recieve." If the pupil is uncertain but happens to guess correctly the first time, it is quite likely that his uncertainty will become apparent on the next presentation. It is true that two correct spellings may have been fortuitous accidents but it is not nearly so likely as in the case of only one correct spelling. Again, let us suppose that a social science teacher wishes to learn whether her pupils know that the Dingley Tariff Law was passed during the presidency of McKinley. She might prepare an item such as the following:

The Dingley Tariff Bill was passed during the presidency of:

- (1) Wilson, (2) McKinley, (3) Harrison, (4) F. D. Roosevelt, (5) Lincoln.

Later in the test an item of this sort might appear:

The tariff bill passed during the administration of McKinley was known as the:

- (1) Wilson Tariff, (2) Payne-Aldrich Tariff, (3) Underwood Tariff, (4) Fordney-McComber Tariff, (5) Dingley Tariff.

If the students answer both of these items correctly, the teacher may be almost certain that they know the approximate time when the Dingley Tariff Bill was adopted. In constructing test items of this sort the teacher must exercise care to avoid giving clues in one of the items that will supply the answers to others.

The teacher who will take the trouble to do so, and every teacher should go through this procedure with some of her tests, will find that the preparation of a chart such as presented on page 146 to show which items were missed

by each pupil is very helpful in determining what sort of remedial teaching is required. It is not suggested that it is worth while to "squeeze" diagnosis from every test but it will be found decidedly worth while to treat the results of some tests in this painstaking way. As previously pointed out, too much of the typical teacher's testing time has been devoted to the giving and scoring of tests and too little time to the critical examination of the results with a view to improving the situation.

From the above discussion it must not be implied that the only instrument of diagnosis is the paper and pencil test. Teachers must be alert to all handicaps to effective school work. The pupil who frequently mispronounces his words in reading or who persists in mis-copying materials from the blackboard may be suffering from defective vision and the pupil who misspells most of his words may be handicapped by defective hearing. Fortunately, simple devices for testing both vision and hearing are finding their way into most schools and in some places nutritional difficulties and other physical handicaps are being detected. When such matters receive consideration complaints concerning pupils' inattention, carelessness, laziness, and indifference usually become less numerous. Where educational diagnosis fails to determine the learning difficulty, a physical diagnosis sometimes will. Naturally, too, the teacher will take into account the pupil's mental capacity as revealed by a valid and reliable test of intelligence, but she must be on guard against any defeatist philosophy that will lead her to say, "Well, this boy has such limited ability, one simply can't expect anything from him."

.

Testing to Motivate Learning

Numerous studies both in the psychological laboratory and in the classroom reveal that learning is facilitated when the results are known to the students. To be most effective, the knowledge of results must come as soon as possible, and must be in specific terms rather than in vague generalities. To say to a pupil that his work in arithmetic is poor is, for example, much less effective than saying that he needs to learn the multiplication combinations. Similarly, to point out that he needs to master 7×9 , 6×7 and 3×8 is still more effective. A pupil may wish to excel in reading or in arithmetic but to do so he must know what it is necessary for him to accomplish. Objectives must, therefore, be clearly defined and should be set at such levels as to be attainable in the not-too-distant future. Thus the pupil who has been helped to set for himself the goal of reading 150 words per minute while reading silently and has had some suggestion about means of checking his progress toward that goal has a much better chance of success than one who is simply encouraged to read better or more rapidly. An appeal to rivalry is often stimulating and may be made by having a school or a class compete against another, by having two groups in the same class compete, or by having the individual compete against his group. Since individuals vary so greatly and since overstimulation and attendant evils may accompany such competition, it is probably better to have the pupil compete against his own previous record. To accomplish this, graphs of various sorts are often effective. In grades above the primary the teacher may well use such terminology as norms, standards, averages, goals, etc., in discussing with the pupils their achievement. Profile charts, by which pupils are able to see graphically how

they stand with respect to their achievement in various school subjects, are sometimes very helpful.

Tests for Evaluating the Teacher's Work

One of the difficulties encountered by supervisors, principals, and superintendents is that of evaluating the work of the individual teacher. Various rating schemes have been devised for this purpose and some attempts have been made to secure objective data for comparing the effectiveness of different teachers in the school system. Information of this sort, if it were highly valid and reliable, would be of inestimable value. The alert administrative official would then know which teachers should be retained, which should be given promotions and better salaries and which should be dismissed. When standardized tests of achievement first became popular, there were several suggestions that it was possible at last to secure this much-desired information. Simply administer a standardized test at the opening of the school year and again at the close of the year and rate the teachers in accordance with the progress shown by the pupils! The scheme seemed absurdly simple but like most simple schemes in the field of education, it was soon shown to have its defects. In the first place, teachers who were aware of the procedure found it profitable to secure the tests in advance and to coach their pupils on the various items. If this were not possible, coaching on similar tests often proved helpful. Since these tests were almost wholly concerned with the acquisition of information or the development of skills, a premium was placed on fact-learning and on spelling and number combination drill. Teachers who were doing a splendid job of developing proper ideals and attitudes and who had attained fine reputations as character build-

ers were suddenly discovered to be less "efficient" under this new method of appraisal than teachers who gave little attention to these matters but devoted their time to drill activities. Fortunately the injustice to teachers was soon discovered by alert administrators and the practice has largely been discontinued. This does not mean, however, that testing at the beginning of the year and again at the end is a practice to be condemned. On the contrary, such testing programs, carefully administered, do give one indication of a teacher's effectiveness. The point is, however, that they give only one index of teaching proficiency. Many other factors must be taken into account in rating teachers.

For the teacher who wishes to determine the most effective method of presenting a given subject, the use of standardized tests is very helpful. Let us suppose, for example, that a teacher wonders whether the teaching of spelling is more effective when the words are presented in context than when taught in columns. By testing she can secure two comparable groups. She can then teach one by the column method and the other by the context method, test again after a period of instruction, and obtain an answer to her question. It is true that teacher-made tests are often useful for such purposes but it will usually be found advisable to make use of standard tests, when they are available, because of the greater care ordinarily exercised in the construction of standardized tests and because of their greater reliability.

The administration of tests at the beginning and at the end of a period of instruction should often be supplemented by testing during the period as well. Such tests will reveal whether the entire group is moving forward or whether there are some pupils who need special attention. As has been pointed out elsewhere, a judicious use of

objective measurement frequently gives the teacher a basis for defense of her methods and an answer to critical attacks by administrators and school patrons.

Co-operative Testing Programs

Co-operative or large-scale testing programs have been initiated in a number of places but for the most part these have been confined to high schools. A study of twenty-six state-wide testing programs indicates that six of them employ intelligence tests or tests of mental ability only; ten employ only achievement tests; and ten employ both types.³ Among those who are friendly to the testing movement there is considerable divergence of opinion concerning the merits of such testing programs. Some persons believe that they have significant uses in supervision, research, motivation, and guidance.⁴ Others feel that the use of achievement tests in such programs is very unfortunate,⁵ especially if the testing is done frequently. One of the dangers in such programs is that they do tend to direct teaching and pupil effort toward those outcomes that have traditionally been measured and hence toward the factual outcomes. There is danger that such testing programs will tend to standardize curricula, a tendency particularly unfortunate in the secondary school where, to a greater extent than in the elementary school, the curricula need to be modified in accordance with the requirements of the community. Other dangers lie in an overemphasis on tra-

³ *Bulletin of Research Activities*, Ohio State Department of Education, Columbus, Ohio, Bulletin R-2, 1938.

⁴ Segel, David, National and State Cooperative High School Testing Programs, *United States Office of Education Bulletin*, No. 9, 1933.

⁵ Douglass, Harl R., The Effects of State and National Testing on the Secondary School, *School Review* 42:497-509, September 1934.

ditional subject matter and the possible discouragement of experimentation. It is a debatable question, but to the writer it seems that if state-wide testing programs were to displace most standardized testing for local purposes, much harm would be done to the development of tests. If schools come to depend upon such programs, it is likely that the only persons vitally interested in developing new and better measuring instruments would be those connected with the central offices from which the tests emanate.

A particularly unfortunate feature of state-wide testing programs, which is not frequently incorporated in them but which has appeared in certain cases, is the contest feature. Schools may be pitted against one another or the individuals making the highest scores in history, English, etc., may then be called to a central place where final tests determine the winner or winners. The attendant excitement would certainly be unfortunate for the elementary pupil and probably does more harm than good to high-school students.

The most common use to which data from these state-wide testing programs has been put is to supply colleges with information concerning the high-school pupils of most promise. Psychological examinations and reading tests are most useful here and little harm can result from a program of this sort. With wholesale graduation from high school there has come to be a need for specific information, particularly for professional colleges and colleges desirous of admitting only a select student body.

Devices for Rapid Scoring

Because of the great amount of time required for scoring the numerous tests given in large school systems or in

state-wide testing programs, considerable demand for mechanical methods of scoring has arisen. One of the developments in this field, and perhaps the most rapid, is the scoring machine produced by the International Business Machines Corporation. Complete scoring with this machine is accomplished at the rate of eight to fourteen papers per minute. Aside from the comparatively high rental costs, which would probably be prohibitive in the typical small school, separate answer sheets would have to be supplied to the pupils. This presents no problem so far as college and high school students are concerned and even pupils in the upper elementary grades accomplish the transfer without difficulty; but just how far down in the elementary grades this device can be used has not yet been adequately determined. An additional feature of machine scoring which appears to the writer to be unfortunate is that only the total score is indicated by the machine and an analysis of individual items requires as much painstaking labor as before. Where tests have been very well developed so that the relative difficulty of the items has been established and where the tests are to be used only for survey purposes, there is little need for item study. Where tests might be used for diagnostic purposes, on the other hand, it is probable that other means of scoring would be more acceptable.

In this connection Stenquist has described a method of scoring which makes use of an obsolescent mimeograph to facilitate checking responses. A separate answer sheet is used by the pupil and these sheets are run through the mimeograph at the rate of 28 to 45 sheets per minute. Thus, for a reasonably short test, the correct answers can be indicated for a group of 45 pupils in about 2 minutes. It then remains for the teacher to count the total number of correct responses but the time required for this is small

in comparison with hand scoring and counting. One of the advantages pointed out by Stenquist is that it permits test scoring while pupil and teacher interest is at its maximum. In the same article Stenquist ⁶ shows a class analysis chart which should be of interest to many readers.

The Self-Marking Tests previously described have the advantage of being convenient for diagnostic use since the correct items are indicated by a cross appearing in a square accompanying the text, or, in the case of longer tests, on the inner pages of the answer booklet. Other devices for rapid scoring are also to be commended for releasing the teacher for other important work. In preparing her own tests, the teacher should give due consideration to problems of scoring lest she may afterwards find them difficult or entirely too time-consuming to score.

Tests and Mental Hygiene

In some instances tests have been known to be the cause of severe mental stress and tension, sometimes even giving rise to fears or anxieties of a serious nature. For this reason it has been claimed that tests injure health. Whether such contentions are justified is not known but many teachers testify that tests do contribute to the nervousness of nervous children. The calm teacher who is careful not to overemphasize the importance of any test, but who leads her pupils to take tests "in stride" as they would any other schoolroom exercise, will find that no undue tension will be created even for the "nervous" child. In connection with this Symonds ⁷ cites a portion of the unpublished

⁶ Stenquist, John L., Devices for Testing, *Nation's Schools* 20:30-33, November 1937.

⁷ Symonds, Percival M., Marks and Examinations as Factors in Personality Adjustment, *National Elementary Principal* 15:355-362, July 1936.

White House Conference on Mental Hygiene in Schools, in which Goodwin Watson points out:

Tests should not be used—

1. To make pupils study something which is not intrinsically motivated
2. As threats, signifying failure, demotion, etc.
3. As climactic events, matters almost of life and death
4. As speed-forcers, suggesting by implication that it matters enormously whether the right answer emerges in half a second or in three seconds
5. As sole bases for evaluating the worth of an individual's contribution, the marks to be given him, the extent of his growth, the nature of his possibilities.

Tests may be used—

1. To help pupils understand how well they do the limited tasks covered by the test
2. To help pupils observe progress in the direction of goals which they have already purposed
3. To assist pupils and teachers in comparing the efficiency of certain methods of study and teaching
4. To suggest weak points needing further educational attention.

PROBLEMS

1. What sorts of graphic devices have you seen employed in schools which serve to motivate pupils' work?
2. Select a problem in methods of teaching which can be solved in part by the use of objective measurement. Describe in detail just how you would go about planning an experiment, what tests you would use, and how you would arrive at a conclusion.
3. Is there a state testing program in your state? What do you know about its merits and demerits? What do you think of such programs in general?

4. What advocates of the formal school or of the informal school be most likely to be in sympathy with ability grouping? Why?
5. Can you recall any situations in which you have seen tests misused? Any situations in which tests were very wisely used? Did the use or misuse appear to reflect the attitude of the teacher toward tests?

BIBLIOGRAPHY

- Beck, Roland L., Progressive Education and Measurement, *Education* 58:557-59, May 1938.
- Broening, Angela M., Tests That Teach, *Baltimore Bulletin of Education* 16:97-106, November-December 1938.
- Corning, Hobart M., *After Testing—What?* Scott, Foresman and Company, 1926.
- Cubberley, E. P., *The Principal and His School*, pp. 358-384, Houghton Mifflin Company, 1923.
- Douglass, Harl R., The Effects of State and National Testing on the Secondary School, *School Review* 42:497-509, September 1934.
- Lefever, D. Welty, What Every Classroom Teacher Should Know about Testing, *Education* 58:520-22, May 1938.
- National Society for the Study of Education, Adapting the Schools to Individual Differences, *Twenty-Fourth Yearbook*, Part II, 1925.
- National Society for the Study of Education, The Grouping of Pupils, *Thirty-Fifth Yearbook*, Part I, 1936.
- Purdom, L. T., *Value of Homogeneous Grouping*, Warwick and York, 1929.
- Terman, L. M., and Others, *Intelligence Tests and School Reorganization*, World Book Company, 1922.
- Torgerson, T. L., What Constitutes an Adequate Testing Program, *Education* 58:553-56, May 1938.
- Wrightstone, J. W., Measuring the Attainment of Newer Educational Objectives, *National Elementary Principal* 16: 493-501.

RECORDING AND REPORTING PROGRESS

Chapter XIV

RECORDING AND REPORTING PROGRESS

AS WE HAVE SHOWN in the previous chapter, not all tests are used for grading purposes. Yet, when tests are mentioned, most persons think of them as instruments for grading rather than as instruments of diagnosis or of instruction. So closely have testing and grading been associated that it is almost impossible to separate the two in the minds of many persons. The fact that the most common use to which test results are still put is that of grading, acts to preserve this association.

In all ages there have probably been those who have rebelled at the notion of assigning marks to indicate the degree of excellence of school work and certainly during the past few years the opposition has been sufficiently vociferous to merit some attention. One of the chief reasons given for the abandonment of school marks is that they are entirely too unreliable. Anyone with even a passing acquaintance with such matters will admit that school marks have been proven in many instances to be very unreliable. As a matter of fact, this point of view as applied to essay examinations has given much impetus to the movement for better and more objective tests.

Uses of School Marks

Among the many uses which have been made of school marks, one may list the following:

1. To provide reports to parents
2. To provide reports to pupils
3. To provide records which aid in the transfer of pupils to other schools
4. To motivate the work of the pupils
5. To indicate the status of the school
6. To aid in determining the time allotment for the various subjects
7. To predict success in higher schools
8. To assist in educational and vocational guidance
9. To classify pupils for instructional purposes
10. To determine which pupils are ready for graduation.

Most persons who have given serious consideration to the matter concede that it is valuable to make some sort of report to parents concerning the progress of their children. It is also rather commonly admitted that the pupils themselves are entitled to know something about the progress they are making. One would scarcely expect a learner to make much improvement in driving a golf ball, for example, unless he could see how far the ball travels or unless he had some way of knowing about the length and direction of its flight. Similarly, nothing is so certain to kill interest in a given school task as continual plodding, day in and day out, without feeling one's progress or achievement. From the point of view of many, therefore, school marks serve *their purpose* mainly as devices for motivation. Considerable opposition to their use in this way has arisen, however, since it is argued that such motivation results in striving for an ulterior goal. It is contended that the pupil should strive to attain mastery for its own sake,

not for the sake of some extrinsic reward; or that the attainment of knowledge should be its own goal and the pupil should have that goal ever before him. Theoretically such objectives appear sound, but it is the experience of most teachers that human frailties thwart them; that these intrinsic values fail to motivate many pupils who can be motivated by school marks.

Opponents of school marks frequently appear to overlook the value of learning motivated by marks, which in turn stimulates interests that might otherwise have lain dormant. It would be difficult to determine how many persons with an initial aversion to a certain branch of science have studied the subject in order to pass an examination in it and who subsequently came to enjoy the subject even to the extent of exploring hitherto unknown areas. Similarly the boy who has studied *Julius Caesar*, only for the sake of bringing home a good report card, may later enjoy a performance of this play infinitely more than if he had not been thus motivated to study it. Unquestionably other devices than marks should be utilized in creating an interest in the school subjects and the subjects themselves must be made as interesting as possible, but it must be realized that some pupils require motivation of an extrinsic nature.

Kinds of School Marks

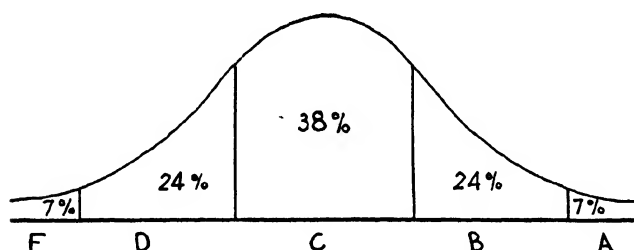
Thirty or forty years ago the most common index of school achievement was the so-called "percentage" grade. Under this system the pupil might receive as much as 100 for "perfect" performance and the "passing" grade in a subject was usually 60, 70, or 75, depending on the pleasure of the local administrator. More recently it has been found that, while there are many different marking sys-

tems, the most common practice is to use some form of letters to designate the different degrees of excellence.¹ Of these, in turn, the most common is a five-point system with the letters A, B, C, and D as the passing grades and E or F indicating failure. Discovery of the unreliability of marks has been the chief contributing factor in providing a five-point grading system rather than a twenty-five- or forty-point system. Thus, since there was ample evidence that teachers could not distinguish accurately enough to assign a reliable mark of 93 to one pupil and 94 to another, it seemed best to give each of these pupils the same grade; namely, A.

Assigning Grades on a Five-Point Scale. Regardless of the type of grading system used, there must be in any given school some agreement as to the relative number of grades of each degree of excellence that are to be assigned. Unless there is such an understanding, one is likely to find that for some instructor the median grade is B or even A, while for another it may be D. Considering that the same pupils are involved, it is extremely unlikely that they should be doing B-grade work in reading, for example, while at the same time doing mostly D-grade work in arithmetic. Having established that the vast majority of traits distribute themselves in accordance with a bell-shaped curve, or normal curve, this fact has often been used in determining the relative number of grades of each degree of excellence that should normally be assigned. If, by way of illustration, the normal curve is divided into five parts by laying off sigma (standard deviation) distances on the base line, we find that the percentage of cases in each division can be computed. In the following

¹ Billett, Roy O., Provisions for Individual Differences, Marking, and Promotions, *United States Office of Education Bulletin* No. 17, 1932.

figure we have marked off one-half sigma distance on each side of the mean, thus embracing about 38 per cent of the cases. An additional sigma distance from each side of the mean will include about 24 per cent of the cases on each side, while most of the remaining 7 per cent on each side will fall within another sigma distance. We may assign 7 per cent A's; 24 per cent B's; 38 per cent C's; 24 per cent D's; and 7 per cent F's in this way.



Some such device is convenient to use when one wishes to convert test scores into letter grades or numerical grades. By way of illustration, the test scores made by thirty pupils on a 130-point test in a seventh grade class in American History are listed below, together with the suggested letter and percentage grades.

<i>Scores</i>	<i>Letter Grades</i>	<i>Percentage Grades</i>
124	A	96
122	A	94
122	A	94
118	B	91
117	B	91
114	B	90
112	B	89
109	B	87
106	C	85
105	C	85

<i>Scores</i>	<i>Letter Grades</i>	<i>Percentage Grades</i>
104	C	84
103	C	83
100	C	82
98	C	81
97	C	80
97	C	80
95	C	79
95	C	79
94	C	79
93	C	78
92	C	78
91	D	77
89	D	76
87	D	75
87	D	75
85	D	74
81	D	72
77	D	70
71	F	68
66	F	64

In assigning grades as is done in this table, the assumption is made that the group is about average in abilities and that the quality of work done in history is fairly typical. In most cases such an assumption will be valid but in some instances the teacher knows that she is dealing with an atypical group and her grading should be modified accordingly. There should be no slavish devotion to the normal curve or to any other grading standard. Teachers are sometimes heard to object to any such standard on the grounds that they do not want to be forced to assign failing grades to a certain number of pupils. This reveals a misunderstanding of such devices since there is nothing about them that demands assignment of failing grades. If one is dealing with a superior group, it may be that no

failing grades should be assigned; if one is dealing with an inferior group, probably A and B grades will be much more rare than is usual. The teacher should have objective evidence from reliable standardized tests to support any contention that a given group is atypical since her subjective judgment in the matter may be quite erroneous.

The above table will serve to call attention to the fact that at times it does not seem advisable to assign the exact percentage of a given letter grade that is called for by the standard employed. For example, we assumed that 7 per cent of the pupils should receive A grades. This would mean two of the thirty pupils. When we find that to follow this rule would lead us to assign A to one pupil with a score of 122 and B to another pupil with the same score, we see that rigid adherence to the plan would be wholly indefensible. We might, therefore, assign only one grade of A or three such grades. Again in making a division between the B and C grades, a question might well be raised. If thirty-one per cent of the grades are to be A's and B's, it would seem that one more pupil should have been assigned a grade of B; namely, the one having a score of 106. This was not done because here there was a more decided "break" in the scores where the division was made. One might, however, easily defend the assignment of a grade of B to the pupil with a score of 106. A somewhat similar situation exists at the point where division between C and D grades is made. When good judgment is coupled with it, the normal curve plan of grading does no one any particular injustice and it does make for a certain amount of uniformity in the assignment of grades.

It will be noted from the above and similar data that a five-point grading system does not relieve one of the necessity of making fine distinctions in grading. It almost never

happens that a sharp distinction is found between the "C" group and the "D" group on one side or the "B" group on the other. Even if one reduces grades to only "satisfactory" and "unsatisfactory" one still has the problem of determining the dividing line between the lowest satisfactory grade and the highest unsatisfactory grade.

The Accomplishment Quotient

One of the oft-repeated criticisms of school marks is that they tend to do an injustice to the pupil who is less favored with the type of mental ability required for successful school work than others. Most persons have known children who, despite diligence and perseverance of a high order, have been consistently assigned failing or inferior grades. It is argued that repeated failures despite tremendous effort result in feelings of inferiority of a very detrimental sort. As a remedy, Franzen² suggested a plan of assigning grades in accordance with the accomplishment quotient (AQ) and this plan was adopted in many schools. The AQ is simply a ratio between mental age and educational age. Thus if a pupil has an educational age, as determined by standardized tests of achievement, of ten years, but a mental age, as found from his performance on an intelligence test, of only eight years, his AQ is 125. Similarly a pupil whose mental age is the same as his educational age would have an AQ of 100. The normal expectancy for any pupil would be an AQ of 100 and any positive deviation would indicate better than average achievement while any figure below 100 would indicate that the pupil was not working up to capacity. Pupils who

² Franzen, R. H., *The Accomplishment Quotient: A School Mark in Terms of Individual Capacity*, *Teachers College Record* XXI: 432-40, November 1920.

because of their chronological ages find themselves in sixth grade when their mental ages are only nine years can hardly be expected to do average work in that grade. Yet we can expect them to do as well as other pupils of the same mental age and if they do so, their AQ's will be 100 which indicate very satisfactory performance.

Because our educational system is so organized as to provide stronger motivation for the duller pupils than for the brighter ones, it was found that the duller pupils in the majority of instances earned the higher AQ's. As a matter of fact, AQ's well above 100 were common for the duller pupils while AQ's well below 100 were the rule for pupils of greater ability. This naturally led to inquiry by the parents but since the system is so easily explained, the parents were usually satisfied with the explanation. When it was desired to report accomplishment in the various school subjects separately, this could be done quite readily by dividing each of the subject ages as determined by standardized tests by the mental age.

The accomplishment quotient technique no longer meets with much favor, though it is still used in a few places. Aside from the fact that its use resulted in assigning the best grades to the dullest pupils in the majority of instances, it lost popularity because of certain scientific studies. McCrory,³ for example, discovered that the AQ had such low reliability as to be quite inadequate for individual diagnosis. Such findings might be expected since none of the standardized achievement tests are perfectly reliable and since also the mental tests are far from being perfectly reliable instruments. Another severe indictment was made by Kelley⁴ who pointed out that although

³ McCrory, J. R., The Reliability of the Accomplishment Quotient, *Journal of Educational Research* XXV:27-29, January 1932.

⁴ Kelley, Truman Lee, *Interpretation of Educational Measurement*, World Book Company, 1927.

achievement tests are designed to test what the pupil has learned in school, whereas intelligence tests strive to measure the child's ability or innate capacity, the two actually have much in common and measure much the same things. From the standpoint of its validity as well as its reliability, it is difficult to defend the use of the accomplishment quotient for grading purposes. In some school systems where its use as a grading device has been discontinued, the AQ is computed for each child because it does yield an additional bit of objective data. Then too it is found to have higher validity in exceptional instances than for the typical case and thus serves to help one understand certain problem situations.

Aside from these statistical objections to the use of the AQ, a larger question is involved. It may well be argued that such a system of grading in no way duplicates the situation which the pupil will find when he leaves school. Certainly it would be the unusual automobile owner who would say, "I take my car to Mr. X for repairs, for while he knows relatively little about motors, he works so diligently and earnestly." Similarly it would be difficult to find a housewife who would say, "I buy my groceries from Mr. Y. He doesn't keep a very good store, but I'm sure he does his very best and so deserves my patronage."

The Meaning of School Marks

The reader has no doubt learned from experience that grades assigned by different teachers and in different schools are based on divergent standards and hence do not mean the same thing. In a school in which considerable attention is paid to a pupil's diligence, a B grade may indicate that the pupil has worked more diligently than the average. In a school in which grades are based on absolute

achievement, such a pupil might receive only a D grade for the same performance. The two chief uses of grades are to inform pupils and parents concerning school progress. Do such grades actually inform? The fact is that they often mislead. Mary brings home a report card which simply indicates a grade of A for each of the subjects she is taking. The parent naturally assumes that Mary is making very satisfactory progress, being quite unaware that the grades are so good because Mary knew all the school was attempting to teach from her previous experiences. Again Mary may be competing with a group of dullards with the result that while her achievement would ordinarily be considered far from satisfactory, she far excels this particular group. The situation would be even more chaotic were it not for the fact that unless there is ability grouping, classes tend to be fairly normal in the distribution of ability and achievement, and that, in general, grades are assigned by comparing a given pupil with the group.

In order to make marks more meaningful, various methods of improving scoring have been suggested. One suggestion is that school marks simply indicate the pupil's present achievement level in each subject. If for reading one found the notation VII, 5, one would know that the pupil was reading as well as the typical seventh grade pupil who had been in that grade for five months. Again the notation VI, 3, in arithmetic would indicate that the pupil had achieved as much in this field as the typical sixth grade pupil who had been in this grade for three months. Such a system of grading has many advantages but if progress reports are to be issued every month or even six weeks, such a system would break down for want of sufficiently reliable and well-standardized tests. It does seem, however, that to give this kind of report to parents and pupils at

the close of each year would be well worth while. At the present stage of test development such reports could undoubtedly be made semi-annually for most school subjects.

How Many Marks Shall There Be?

One of the obvious difficulties with school marks has been that they are expected to represent all that the teacher has observed about the pupil, with the possible exception of his "deportment," whatever that term may mean. A recent commendable tendency is that of entering on the pupil's record data on many things other than his academic achievement. As Goodwin Watson ⁵ says:

Marks would be improved by giving more of them. Existing reports seldom give a fair picture of individual achievement. Many facts about John are much more important than an "average" of B in arithmetic. There is the fact that he stayed after school to help the teacher clean out some cupboards. There is the fact that he is the best baseball pitcher in his grade. There is the fact that he went thru a period of confusion about the figuring of discounts on notes, tried a little, gave up, got angry and disgusted with the whole business, was led by a calm teacher to make another trial, found success increasing, and has now mastered that unit of work. There is the fact that he is jealous of the position achieved by the boy who was elected class president and so has made several slighting remarks about that boy. There is the fact that he got into a fight to avenge a smaller boy who was being taunted about his nationality, the same nationality as John's. There is the fact that he enjoyed a Victrola rendition of the Pilgrim's Chorus intensely, and was surprised to find that he liked music so well.

⁵ Quoted by Symonds, Percival M., in *Marks and Examinations as Factors in Personality Adjustment*, *National Elementary Principal* 15:355-363, July 1936.

To meet the need for more marks, we find report cards appearing with evaluations of such traits as:

- Ability to work and play with others
- Self-control
- Health habits
- Physical condition
- Courtesy
- Initiative
- Work and study habits
- Use of free time
- Motor co-ordinations
- Ability to construct

Those who object to such ratings argue with good reason that teachers' estimates of these characteristics are unreliable. It may be noted, however, that the teacher who must observe these traits in order to rate them on the child's report will probably improve in rating as she gains experience. Certainly her attention will be directed periodically to these traits if she has failed to pay much attention to them formerly. In this connection, the writer wishes to call attention to the fact that a five-point and particularly a two-point grading system for achievement can hardly be defended at the present time. It may have been true at one time that most teachers could not make finer judgments of pupil progress. With the more abundant use of reliable tests many teachers are well able to distinguish grades of C+ or C- from grades of C, and probably many could use grades of B-, B, and B+ and D-, D, and D+. Every possible opportunity should be provided for them to make more discriminating judgments and to develop tests that will help them toward this end. The time has probably come for the adoption of 7-point, 9-point, or 11-point grading systems. Certainly the

tendency to use a two-point system appears to be in exactly the wrong direction.

A Good Marking System

By way of recapitulation of some of the previous discussion we may point out the following desirable characteristics of school marks:

1. The marks should be as reliable as possible. This means that they must be based on reliable measures.

2. Marks should be as discriminating as possible so as to show whatever progress has been made.

3. Marks of progress in school subjects should be based principally on objective tests lest too many subjective and irrelevant elements enter. This means that schools should abandon the plan used in some places of permitting pupils with certain scholastic averages to be exempt from final examinations. This is, in any case, a doubtful procedure since it tends to give the pupils the notion that tests are instruments of torture from which they may be privileged to escape. Furthermore the pupil who has had little experience with tests may find himself at a disadvantage when he is promoted or transferred to a school in which examinations are required of all.

4. Marks of progress in school subjects should be based simply on achievement and scholarship. Parents should also know how well the pupil is working, but a separate index of this trait should be used.

5. Marks of the various teachers and, if possible, of various schools, should be comparable.

6. Records, at least those sent to parents, should indicate other things than progress in school subjects.

Records of School Progress

If maximum value is to be obtained from testing programs, adequate records must be made and kept for refer-

ence from year to year. For this purpose the time-honored school register is neither adequate nor convenient, although the school codes in many states still require their use. It is not adequate because, as a rule, insufficient space is allotted for each pupil so that one cannot record all the desired items, such as test scores. It is not convenient because if one were to bring each pupil's record up to date, it would necessitate re-copying a great many data. If this is not done, one must look back through a series of registers to obtain an adequate picture of the pupil's progress. As a result, the typical teacher is likely to neglect this source of information despite the fact that she might derive much help from it in understanding her pupils.

The more recent tendency in connection with school records has been to utilize some kind of card system. Cards are much more convenient and can be arranged in such a way as to provide for the recording of all needed data. The chief difficulty with a system of this kind is that an individual's card may become lost, particularly if the records are kept in the teacher's room where they are most accessible to her and hence most useful. To guard against loss, the records are frequently kept in the principal's office or in a central records office, where they are, of course, less accessible to the teacher. One of the best solutions to this problem is to keep the most comprehensive record in the classroom (or the teacher's office, if she has one), and to transfer from this at stated intervals the most pertinent information to a card kept in the central office.

Among the characteristics of an adequate record system, one might name the following as being of special importance:

1. Adequate space should be provided for the retention of data that will continue year after year to

be relevant and useful. At the same time, such records should not be encumbered with irrelevant data which become meaningless in a short time and add greatly to the labor of record-keeping. Records of performance on standardized tests are almost sure to be informative. Whether records of teacher-made tests should appear on the cumulative record is a debatable point. If they do appear, care should be taken to make them meaningful by recording them in terms of the grading system employed by the school.

2. The records should be cumulative. In most school systems it will be found advantageous to provide space for pertinent data for each grade of the elementary and high school.

3. The forms should necessitate a minimum of repetition. Many teachers consider record-keeping as a form of drudgery and they are likely to rebel if much of the labor is concerned with copying from one form to another. A large permanent card makes it unnecessary to copy the child's name and previous educational, medical, and social history.

4. The system should be compact and accessible.

5. The forms used should be sufficiently durable to withstand years of use by different persons.

6. If transfers are to be made from one form to another, provision for easy transfer should be made.

PROBLEMS

1. Indicate how school marks might be used in determining the time allotment for the various school subjects.
2. Using the normal curve concept as a general guide, would you assign grades just as was done on page 327? If not, what variations would you make and how would you justify them?
3. Do you agree with the statement "our educational system is so organized as to provide more motivation for duller pupils than for brighter ones"? If so, what is the motivating influence? Could you organize a school in which all pupils would be motivated to the same degree?

4. Do you agree that considering the present stage of development in tests it would be unwise to make a monthly report based solely on the grade equivalent of scores made on standardized tests? Why?
5. If possible, secure from a near-by school system a sample of the pupil-record form used. Evaluate it in accordance with the standards enumerated for a good record system.

BIBLIOGRAPHY

- Adams, A. Elwood, Marking Pupils on Their Working Ability, *Nation's Schools* 17:35-6, April 1936.
- Clark, Ridgley C., The Status of the School Mark, *American School Board Journal* 94:39-40, March 1937.
- Dawson, Mildred A., Report Cards Without Marks, *Journal of Education* 118:532-4, December 2, 1935.
- Harrington, Don, Sensible Grading System, *Nation's Schools* 21:37-8, February 1938.
- Hauser, L. J., We Shift to New Report Cards, *Journal of Education* 119:68-9, February 3, 1936.
- Hill, George E., The Report Card in Present Practice, *Educational Method* 15:115-131, December 1935.
- Hobbs, Valine, A Report-Card Experiment, *Instructor* 45:20 and 75, September 1936.
- Kelley, Truman Lee, *Interpretation of Educational Measurement*, World Book Company, 1927.
- Lafferty, H. M., Poor Pupils or Poor Marking Systems, *School Executive* 56:410-11, June 1937.
- Lindquist, E. F., Changing Values in Educational Measurement, *Educational Record*, Supplement to Volume 17: 64-81, October 1936.
- Marple, C. H., A Viewpoint on "School Marks," *American School Board Journal* 95:31-34, July 1937.
- McCrory, J. R., The Reliability of the Accomplishment Quotient, *Journal of Educational Research* XXV:27-29, January 1932.
- Moore, C. C., Needed: A New Plan of Grading, *School Executive* 56:402-3, June 1937.
- Owen, W. B., Making the Grade, *Nation's Schools* 18:21-22, August 1936.
- Segel, David, To Mark or Not to Mark—An Unsolved Problem, *School Life* 22:34, October 1936.
- Smith, Myron E., Why Mark? *Journal of Education* 119:355-8, September 7, 1936.

APPENDIX

Publishers and Distributors of Tests Mentioned in This Volume

1. Bureau of Educational Measurements and Standards, Kansas State Teachers College, Emporia.
2. Bureau of Educational Research, University of Illinois, Urbana.
3. Bureau of Educational Research and Service, University of Iowa, Iowa City.
4. Courtis Standard Tests, Detroit, Michigan.
5. Educational and Personnel Publishing Company, Washington.
6. Educational Test Bureau, Minneapolis and Philadelphia.
7. Farnum Press, Minneapolis.
8. C. A. Gregory, Inc., Cincinnati, Ohio.
9. E. M. Hale and Company, Milwaukee, Wisconsin.
10. Houghton Mifflin Company, Boston.
11. J. B. Lippincott Company, Philadelphia.
12. Lyons and Carnahan, Chicago.
13. Public School Publishing Company, Bloomington, Illinois.
14. Russell Sage Foundation, New York.
15. Scott, Foresman and Company, Chicago.
16. Southern California School Book Depository, Los Angeles.
17. Teachers College Bureau of Publications, Columbia University, New York.
18. University of Texas, Austin.
19. World Book Company, Yonkers-on-Hudson, New York.



INDEX

INDEX

- Ability grouping, 304
 - arguments for and against, 306
- Accomplishment quotient, 330
- Achievement tests, 40
- Adams, A. Elwood, 340
- Administration of tests, 279
- Age norms, 286
- Analytical tests, 42
- Andruss, Harvey A., 73, 79
- Aptitude tests, 38-40
- Arithmetic, problems of meas-
urement in, 120
 - survey tests in, 123
 - diagnostic tests in, 125
- Art, tests in, 195
- Ashbaugh, E. J., 61, 158
- Ayer, Adelaide, 182
- Ayres Handwriting Scale, 151
- Ayres Spelling Scale, 157

- Beach Music Test, 193
- Beck, Roland L., 319
- Betts, E. A., 134
- Bevins, Alice E., 200
- Bibliography, 33, 56, 79, 105,
134, 163, 182, 200, 226, 247,
297, 340
- Billett, Roy O., 326
- Binet-Simon Test, 253 ff.
- Bobbitt, Franklin, 205
- Boynton, Paul L., 271
- Breed, F. S., 155, 164
- Briggs and Armacost, 79
- Broening, Angela M., 319
- Brooks, Fowler D., 196, 200
- Brown, Ralph R., 253, 271
- Brueckner, L. J., 44, 135
- Brueckner and Melby, 41, 56,
135, 163, 182
- Buckingham-Ayres Spelling Scale,
157
- Buckingham-Stevenson Place Ge-
ography Tests, 178
- Burton Civics Test, 174
- Buswell-John Diagnostic Chart
in Arithmetic, 125
 - Teaching and Practice Exer-
cises in Arithmetic, 130

- Cajori, M. H., 182
- California Test of Mental Ma-
turity, 259
- Cattell, Psyche, 254
- Cause-Effect Test, 215
- Central tendency, measures of,
85
- Chance factors in tests, 75
- Character Education Inquiry,
223
- Character Tests, 214
- Characteristics of formal tests, 25
- Charles, J. W., 74
- Charters Diagnostic Language
Test, 144
- Chenoweth and Selkirk, 226
- Child Development Readers, 119
- Chipman, C. E., 253, 271
- Civics Tests, 174
- Civilization and measurement,
relation between, 16
- Clapp Drill Books in Arithmetic,
130
- Clapp-Young Arithmetic Test,
123

- Clapp-Young English Test, 144
 Clark Letter Writing Test, 149
 Clark, Ridgley C., 340
 Clark-Otis-Hatton Instructional Tests in Arithmetic for Beginners, 131
 Class, E. C., 79
 Classification of tests, 37
 Cole, Luella, 163, 182
 Cole and von Borgrersrode, 294
 Compass tests in arithmetic, 124, 127
 Composition scales, 146
 Constructing Tests, 59
 Cook, W. W., 156, 164
 Co-operative Testing Programs, 313
 Corning, Hobart M., 319
 Correcting for chance, 75
 Correlation coefficient, determination of, 95 ff.
 Curtis Standard Research Test in Music, 194
 Criteria for selecting tests, 275 ff.
 Cubberley, E. P., 319
 Curriculum tests, 44
 Curtis, F. D., 213
- Dawson, Mildred A., 340
 Dearborn, Walter F., 134
 Deffenbaugh, W. S., 269, 271
 DeGraff and Ruch, 74
 DeMay-McCall Standard Test Lessons in Fractions, 131
 Denny, E. C., 158
 Denny-Nelson Tests in American History, 42, 173
 Development of testing, 30
 Diagnostic tests, 43, 307
 in arithmetic, 125
 preparation of, 307
 Donnelly, Helen E., 134
 Douglas and Holland, 271
 Douglass, Harl R., 313, 319
- Drawing scales, 195
 Durrell, Donald D., 134
 Duties test, 217
- Economy Remedial Exercises, 130
 Eng, Helga, 200
 Enlow, E. R., 105
 Essay test, 45, 60
 Evaluating work of teachers, 311
- Fitzpatrick, F. L., 226
 Formal and informal tests, 75
 Forms of tests, 45
 Forsythe and Rugen, 226
 Franseen Diagnostic Tests in English, 144
 Franzen, R. H., 330
 Frazier, C. F., 157
 Freeman Chart for Diagnosing Faults in Handwriting, 153
 Freeman and Dougherty, 163
 Frequency table, 84
 Frutchey, F. P., 79
 Fullerton, C. A., 200
 Function of tests, 38
- Garrett, H. E., 105
 Gates, A. I., 112, 114, 134
 Gates Graded Word Pronunciation Test, 114
 Gates Primary Reading Tests, 110
 Gates Silent Reading Tests, 43, 111
 Gates-Strang Health Knowledge Test, 207
 General tests of achievement, 231
 Geography:
 objectives of teaching, 175 ff.
 tests in, 178
 Gildersleeve Musical Achievement Tests, 193

- Gilliland, Jordan, and Freeman, 247
- Good, Warren R., 105
- Goodenough, Florence L., 200
- Grade norms, 285
- Grading, 323 ff.
on five-point scale, 326
on percentage scale, 328
- Grammar tests, 145
- Gray, C. T., 152, 163
- Gray's Oral Reading Check Tests, 113
- Greene, H. A., 141, 163
- Greene and Jorgensen, 56, 79, 134, 135, 163, 247, 298
- Gregory-Hagerty Geography Tests, 178
- Guessing in tests, 75
- Guiler, W. S., 156
- Guiler Diagnostic Tests in Punctuation and Remedial Exercises, 150
- Haggerty Intelligence Examination, 256
- Handwriting, tests in, 150 ff.
problems of measurement in, 154
- Harrington, Don, 340
- Hartshorne and May, 215 ff.
- Hartshorne, May, and Maller, 226
- Hartshorne, May, and Shuttleworth, 226
- Hauser, L. J., 340
- Hawkes, Lindquist, and Mann, 33, 49, 79
- Health education, 205
- Henmon, V. A. C., 44, 56
- Henmon-Nelson Tests, 39, 259
- Hevner, Kate, 79
- Hill, George E., 340
- Hill Test in Civic Attitudes, 168
- Hillbrand Sight-Singing Test, 193
- History, tests in, 173
objectives of teaching, 171
- Hobbs, Valine, 340
- Holzinger, Karl J., 105
- Homogeneous grouping, 304
arguments for and against, 306
- Horn, Ernest, 164, 171
- Horn, Helen R., 79
- Horns, J. W., 197
- Hutchinson Music Test, 193
- Importance of measurement, 15
- Instructional Reading Tests for Intermediate Grades, 111
- Intelligence, definitions of, 251
controversies concerning, 251
maturing of, 252
nomenclature, 263
quotient, 262
relation to school progress, 264
- Intelligence tests, 19
individual, 253 ff.
group, 256
kindergarten-primary, 260
non-language, 261
uses of, 266
- Iowa Reading Test, 111
- Iowa Spelling Scales, 158
- Irrelevant Terms items, 54
- Jacobson and Van Dusen, 115
- Jersild and Bienstock, 200
- Johnson, Ruth, 200
- Jones, J. H., 226
- Kelley, Truman Lee, 298, 331, 340
- Kinesthetic Method in Reading, 117
- Kinter and Achilles, 200
- Kline-Carey Measuring Scale for Freehand Drawing, 196

- Kramer, Edna E., 105
 Kuhlmann-Anderson Tests, 258
 Kwalwasser-Dykema Music Tests, 189
 Kwalwasser-Ruch Test of Musical Accomplishment, 192

 Lacy, Frances, 226
 Lafferty, H. M., 340
 Lang, A. R., 33
 Language, tests in, 139
 uses of, 139, 145
 practice materials, 149
 Language usage tests, 142
 Lefever, D. Welty, 319
 Letter-writing scales, 148
 Lewerenz, Alfred S., 201
 Lewerenz Tests in Fundamental Abilities of Visual Arts, 196
 Lindquist, E. F., 105, 340
 Lindquist and Anderson, 182
 Literature tests, 150
 Lyman, R. L., 163
 Lu Pone, O. J., 226

 Marking systems, desirable characteristics of, 336
 Marple, C. H., 340
 Matching exercises, 53
 constructing, 71 ff.
 McAdory, Margaret, 201
 McClusky, H. Y., 79
 McCrory, J. R., 331, 340
 McKee, Paul, 163, 164
 Mean, arithmetic, 87
 Measurement of quality, 20, 23
 Measurement of quantity, 19, 23
 Measurement in daily life, 15
 Measurement in education, 23
 nature of, 23
 present status, 29
 beginnings, 30
 Median, 86 ✓
 Mental age, 262
 Mental hygiene and testing, 316
 Merton, E., 120
 Metropolitan Achievement Tests, 241
 in spelling, 160
 Mode, 85
 Modern School Achievement Test, 235
 in spelling, 160
 in science, 212
 Monroe and Streitz, 147
 Monroe Reading Aptitude Tests, 119
 Monroe Silent Reading Test, 110, 118
 Moore, C. C., 340
 Morrison-McCall Spelling Scales, 160
 Morton, R. L., 135
 Motivation of learning, 310
 Mullen and Lanz Exercises and Tests in English, 150
 Multiple Choice items:
 single response, 51
 plural response, 52
 constructing, 66 ff.
 Mursell, James L., 200
 Music, objectives of teaching, 185
 tests in, 186
 use of tests in, 195
 Musical appreciation, tests of, 194
 Musical talent, tests in, 187

 Nelson, M. J., 304
 Nelson and Denny, 105, 156, 164
 Nelson Silent Reading Test, 43, 86, 87, 111
 Normal Curve, 327
 Norms, 283
 adequacy of, 248
 Norton and Norton, 226

- Objective tests, types of, 49
- Odell, C. W., 33, 298
- Opposition to tests, 301
- Oral English testing, 141
- Oral presentation of tests, 76
- Oral reading tests, 113
- Orleans, Jacob S., 33
- Orleans-Solomon Prognosis Test, 40
- Otis Classification Test, 41, 232
- Otis Self-Administering Tests of Mental Ability, 256
- Owen, W. B., 340

- Patterson, S. W., 134
- Pearson Product Moment Method of Correlation, 94 ff.
- Percentile scores, 287
- Personality tests, 214
- Pintner, R., 271
- Pintner Non-language Tests, 261
- Pintner, Rinsland, and Zubin, 156
- Practice materials in arithmetic, 130
- Pressey, L. C., 182
- Pressey and Pressey, 115, 182
- Pressey-Richards Tests of Understanding of American History, 173
- Pre-tests, 303
- Pribble-Brezler Exercises in English, 150
- Pribble-McCrory Diagnostic Tests in Practical English Grammar, 143
- Probability Test, 217
- Probable error of a score, 291
- Problems, 32, 54, 77, 101-104, 133, 161, 181, 199, 225, 246, 268, 297, 317, 338
- Prognostic tests in music, 187
- Progressive achievement tests, 242
- Provocations test, 218
- Public School Achievement Tests:
 - in History, 173
 - in Health, 211
- Purdum, L. T., 319
- Purpose of testing, 275

- Quartile points, 90 ff.

- Range, 90
- Rating scales for tests, 293
- Rating teachers, 311
- Reading:
 - importance of, 109
 - improvement of, 114
 - kinesthetic method in, 117
 - readiness tests, 119
 - teacher-made tests in, 117
 - tests in, oral, 113
 - tests in, silent, 110
- Recall-completion items, 52
 - constructing, 70 ff.
- Records, school, 323
- Relationship, measures of, 94
- Reliability of tests, 49, 289
 - factors affecting, 292
- Renfrow Sixth Grade History Tests, 174
- Rice, J. M., 30
- Riebe, Nelson, and Kittrell, 306
- Rinsland, H. D., 79
- Robertson, Martin L., 226
- Rogers, F. R., 227
- Ruch, G. M., 56, 61, 75, 79
- Rugg, Harold O., 172

- Sampling, 48
- Samuelson, Agnes, 200
- Sangren, Paul V., 134
- Sangren Information Tests for Young Children, 245
- Sangren-Woody Reading Tests, 43, 111

- Scattergram for determining correlation, 97
- Scheidemann, Norma V., 69, 79
- Schoen Music Tests, 191
- School marks, 323 ff.
 kinds of, 325
 meaning of, 332
 opposition to, 325
 uses of, 324
- School progress and intelligence, 264
- School records, 336
 characteristics of good, 337
- Science, elementary, 211
 tests in, 212
 objectives in, 213
- Scoring of tests, 280
 devices for, 314
- Seashore Musical Talent Tests, 40, 187
- Segel, David, 313, 340
- Selection of tests, criteria for, 275 ff.
- Self-Marking Achievement Test, 244
 in spelling, 161
- Selke, Erich, 157, 164
- Semi-inter-quartile range, 90
- Short-cut method:
 for determining mean, 88
 for determining standard deviation, 93
- Sigma, 92
- Silent reading tests, 110
- Silver, Harry B., 227
- Sims and Knox, 80
- Smith, J. Russell, 175, 179
- Smith, Myron E., 340
- Smith and Wright, 56, 134, 135, 163, 182, 298
- Social studies:
 problems of measurement, 167
 tests in, 167
 integration of, 171
- Social studies (Cont.):
 objectives of teaching, 167, 176
 teacher-made tests in, 179
 uses of tests in, 180
- Sorenson, Herbert, 105
- Spearman-Brown formula, 290
- Special aptitude tests, 39
- Spelling:
 problem of measurement in, 155
 multiple choice test in, 156
 lists, 157
- Standard deviation, 92
- Standardized tests, compared with yardstick, 27
- Stanford Achievement Tests, 41, 233
 in reading, 110
 in spelling, 159
- Stanton, H., 200
- Starch, Daniel, 154, 163
- Starch and Elliott, 47
- State-wide testing programs, 313
- Statistical Procedures, 83 ✓
- Stenquist, John L., 316
- Stenquist Mechanical Aptitude Tests, 40
- Stone, Clarence R., 115, 164
- Sullivan, Helen Blair, 134
- Survey tests in arithmetic, 123
- Symonds, Percival M., 316
- Symonds and Chase, 150
- Systems of measurement, 17
- T-scores, 288
- Tabulations of data, 83
- Teacher-made tests, 59 ff.
 in reading, 117
 in arithmetic, 131
 in spelling, 157
 in language, 146
 in social studies, 179
- Teacher rating, 268

- Terman, L. M., 254, 262, 264, 271, 319
- Terman and Merrill, 254, 262, 271
- Test-rating scales, 293
- Thorndike, E. L., 154, 164
- Thorndike's Scale for General Merit of Children's Drawings, 195
- Thralls, Zoe, 176
- Tiegs, E. W., 33, 56, 80, 134, 135, 152, 163, 164, 298
- Torgerson, T. L., 224, 319
- Torgerson and Matthies, 129
- Traxler, Arthur E., 227
- Trow, William Clark, 263, 271
- True-false statements, 50
constructing, 63 ff.
- Typography of tests, 282
- Uhl-Hotz Practice Lessons in English, 150
- Unit Scales of Attainment, 238
in spelling, 160
in science, 212
- Upton-Chassell Citizenship Scales, 174
- Uses of tests, 301
for diagnosis, 307
for grouping, 304
in arithmetic, 130
in art, 197
in language, 145
in music, 195
- Uses of tests (Cont.)
in reading, 114
in social studies, 180
of intelligence, 266
of pre-tests, 303
- Vacations, influence of, 304
- Validity, curricular, 276
statistical, 278
- Van Wagenen American History Scales, 169
- Variability, measures of, 90
- Webb and Shotwell, 185, 247
- Wert, James E., 105
- West, P. V., 164
- Whalen, Mary A., 200
- Whitford, W. G., 201
- Wiedefeld-Walther Geography Test, 42, 178
- Willing Composition Scale, 148
- Wilson, Guy M., 160
- Wilson and Hoke, 235, 247
- Wilson Language Error Test, 144
- Witty, Paul A., 164
- Woody-McCall Mixed Fundamentals in Arithmetic Scales, 42
- Wrightstone, J. W., 319
- Yes-no items, 51
- Zero point in measurement, 20

72 $\overline{) 10.06}$
 $\underline{- 72}$
 286
 $\underline{- 286}$
 0



125 456

UNIVERS
LIBRARY

